STUDIA INFORMATICA 2011
Volume 32 Number 2B (97)

Piotr BAJERSKI, Tomasz PŁUCIENNIK
Politechnika Śląska, Instytut Informatyki

SPECYFIKACJA DANYCH INSPIRE DLA NAZW GEOGRAFICZNYCH I UWARUNKOWANIA EFEKTYWNOŚCI JEJ IMPLEMENTACJI

Streszczenie. Niniejszy artykuł prezentuje model danych INSPIRE dla nazw geograficznych, wymagania dotyczące ich wyszukiwania oraz omawia uwarunkowania efektywności ich implementacji w bazach danych. Przedstawiono porównanie funkcjonalności wiodących systemów zarządzania bazami danych w kontekście realizacji zadań wyszukiwania, zawierających predykaty na nazwach geograficznych.

Słowa kluczowe: bazy danych przestrzennych, Dyrektywa INSPIRE, multimedialne bazy danych, nazwy geograficzne

INSPIRE DATA SPECIFICATION ON GEOGRAPHICAL NAMES AND DETERMINANTS OF ITS IMPLEMENTATION EFFICIENCY

Summary. This article presents the INSPIRE data model for geographical names, requirements regarding searching for geographical names and discusses the determinants of their efficiency in databases implementations. Comparison of functionalities of the leading database management systems in the context of implementation of search tasks containing predicates for geographical names is presented.

Keywords: geographical names, INSPIRE Directive, multimedia databases, spatial databases

1. Wstęp

Od 15 maja 2007 roku obowiązuje dyrektywa Parlamentu Europejskiego i Rady nr 2007/2/WE z dnia 14 marca 2007, ustanawiająca infrastrukturę informacji przestrzennej we Wspólnocie Europejskiej (ang. *Infrastructure for Spatial Information in the European*

Community, INSPIRE) [2], nazywana w skrócie Dyrektywą INSPIRE. Dyrektywa ta została transponowana do polskiego prawa Ustawą o infrastrukturze informacji przestrzennej z dnia 4 marca 2010.

Głównymi elementami infrastruktury informacji przestrzennej, budowanej w ramach realizacji dyrektywy INSPIRE, są: dane przestrzenne, opisujące je metadane oraz usługi służące do ich wyszukiwania, przeglądania, pobierania, przekształcania i koordynacji pracy innych usług danych przestrzennych. Dyrektywie INSPIRE towarzyszą rozporządzenia Komisji Europejskiej, wytyczne techniczne oraz opracowania techniczne. W aneksach Dyrektywy INSPIRE została przedstawiona klasyfikacja danych przestrzennych. Aneks I wymienia tematy, które zostały uznane za najważniejsze dla infrastruktury i które mają zostać zharmonizowane w najbliższych latach: systemy odniesienia za pomocą współrzędnych, systemy siatek geograficznych, nazwy geograficzne, jednostki administracyjne, adresy, działki katastralne, sieci transportowe, hydrografia i obszary chronione. W zakresie całej infrastruktury informacji przestrzennej ważną rolę odgrywają nazwy geograficzne, będące jednymi z podstawowych typów danych referencyjnych, tzn. danych tworzących w ogólności przestrzenne ramy określania położenia geograficznego oraz w szczególności pozwalających na tworzenie powiązań i/lub wskazywanie na inne informacje należące do specyficznych zakresów tematycznych, takich jak: środowisko naturalne, adresy, zarządzanie przestrzenią, zdrowie ludności itd. [4].

W zakresie specyfikacji danych INSPIRE dostępne są następujące zasoby:

- ogólny model koncepcyjny [4],
- specyfikacje danych dla poszczególnych tematów z aneksów Dyrektywy INSPIRE,
- reguły zapisu danych w języku XML,
- skonsolidowany model INSPIRE w języku UML (ang. *INSPIRE Consolidated UML Model*),
- schematy aplikacyjne w języku GML (ang. *GML Application Schemas*) dla poszczególnych tematów z aneksów Dyrektywy,
- rejestr list kodowych INSPIRE (ang. INSPIRE *Code List Dictionaries*).

Modele INSPIRE bazują na modelach ISO serii 19100 oraz OGC. Wszystkie dokumenty prawne INSPIRE można pobrać bezpłatnie z serwisu http://eur-lex.europa.eu, a dokumenty techniczne ze strony http://inspire.jrc.ec.europa.eu/.

Modelowi danych nazw geograficznych poświęcono rozdział 2. W rozdziale 3. przedstawiono wymagania względem wyszukiwania po nazwach geograficznych, a w rozdziale 4. przedstawiono analizę możliwości spełnienia tych wymagań przez dostępne systemy zarządzania bazami danych. W rozdziale 5. zaproponowano metody optymalizacji zapytań zawierających predykaty na nazwach geograficznych. Artykuł kończy podsumowanie.

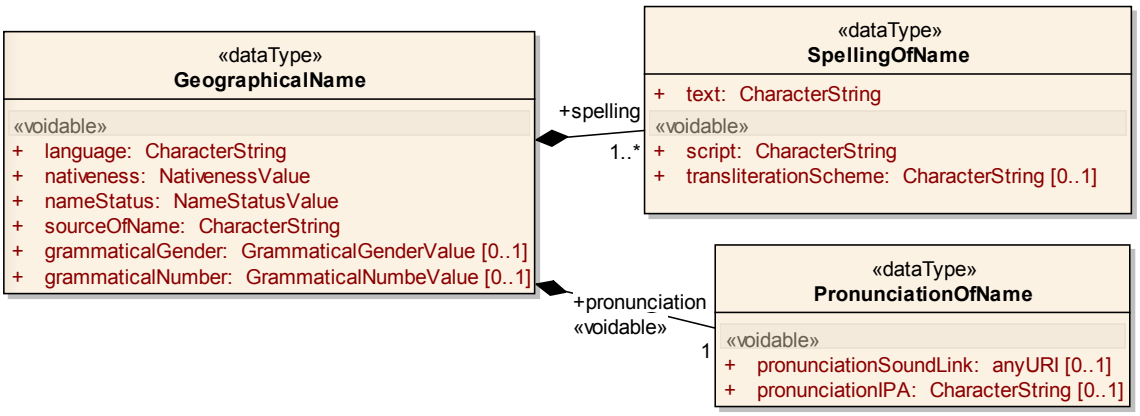
2. Modele danych dla nazw geograficznych

Zakres informacyjny nazw geograficznych został zdefiniowany w rozporządzeniu Komisji (UE) nr 1089/2010 z dnia 23 listopada 2010 r. w sprawie wykonania dyrektywy 2007/2/WE Parlamentu Europejskiego i Rady w zakresie interoperacyjności zbiorów i usług danych przestrzennych [3]. Model UML oraz zapis nazw geograficznych przedstawiono w wytycznych technicznych INSPIRE *Data Specification on Geographical Names – Guidelines* [5].

Podczas opracowywania modelu danych INSPIRE dla nazw geograficznych wykorzystano wcześniejsze doświadczenia organizacji zajmujących się standaryzacją i udostępnianiem nazw geograficznych, takich jak UNGEGN (*United Nations Group of Experts on Geographical Names*) i EuroGeoNames, oraz opracowujących modele danych przestrzennych (OGC i komitet ISO/TC211). W normach serii ISO 19100 i specyfikacjach OGC brak jest modelu danych dla nazw geograficznych. Pojęciowo najbliższy jest model gazetera (ang. *gazetteer*), zawarty w normie ISO 19112 (*Geographic Information – Spatial Referencing by Geographic Identifiers*), który w ramach modelu danych INSPIRE został zmodyfikowany. Znaczące jest, że w rozporządzeniu w sprawie zapewnienia interoperacyjności zbiorów i usług danych przestrzennych [3] nie podano wymagań względem gazetera, określono jednak wymagania dotyczące nazw geograficznych.

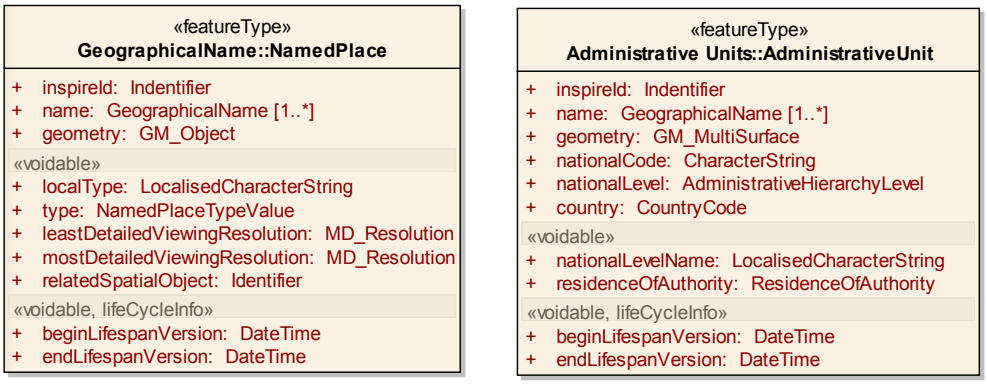
W modelu danych INSPIRE nazwa geograficzna jest modelowana przez klasę *GeographicalName* (rys. 1), przechowującą właściwy sposób lub sposoby zapisania nazwy encji przestrzennej (*spelling*) w języku określonym kodem z normy ISO 639-3 lub z ISO 639-5 (*language*). Każda nazwa posiada najwyżej jeden status (*nameStatus*) i jedno źródło danych (atrybut *sourceOfName*). Można również określić klasę rzeczowników (*grammaticalGender*) oraz kategorię gramatyczną rzeczowników, wyrażającą różnice w liczbie (*grammaticalNumber*). Ze sposobem zapisu (*text*) można zapamiętać zastosowany zbiór symboli graficznych (*script*) oraz metodę konwersji nazw między różnymi sposobami zapisu (*transliterationScheme*). Przechowywanie wielu sposobów zapisu jest stosowane w przypadku, gdy istnieje kilka poprawnych sposobów zapisu danej nazwy w określonym języku. Atrybut *nativeness*, którego dziedziną jest lista kodowa o wartościach *endonym* i *exonym*, określa odpowiednio, czy nazwa jest zapisana w języku używanym na obszarze używania języka, w którym jest zapisany, czy poza nim. Przykładowo, chcąc zapisać nazwę miasta Kraków w językach polskim i angielskim należy utworzyć dla encji przestrzennej miasto Kraków dwie nazwy geograficzne: pierwszą z wartościami atrybutów: *languge* = „pol”, *nativeness = „endonym”* i *text* = „Kraków” i drugą z wartościami odpowiednio: *languge* = „eng”, *nativeness = „exonym”* i *text* = „Cracow”.

Uwagę zwraca występowanie tylko jednego nagrania wymowy dla nazwy geograficznej, chociaż może być podanych wiele zapisów tej nazwy. W przypadku gdy istnieje więcej niż jedna wymowa, należy utworzyć dodatkową nazwę, nawet jeżeli zapisy są takie same.



Rys. 1. Uproszczony diagram klas, modelujący nazwy geograficzne
Fig. 1. Simplified class diagram modeling geographical names

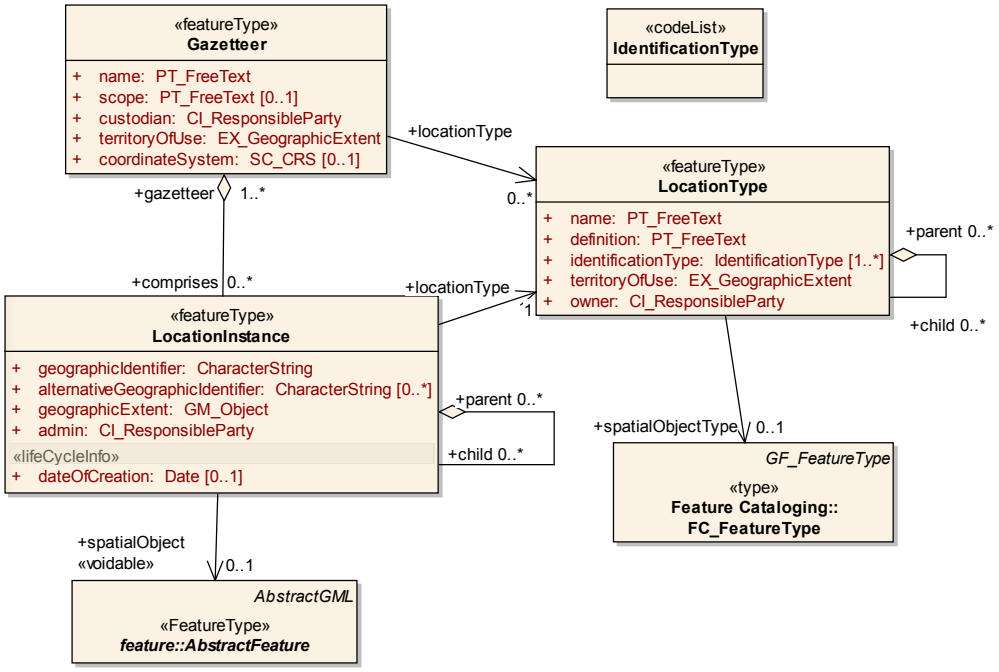
W celu ilustracji wykorzystania nazw geograficznych, na rys. 2 przedstawiono ich użycie w definicji typu przestrzennego *NamedPlace* (*nazwane miejsce)* i w modelu danych jednostek administracyjnych. Stereotyp *<<featureType>>* oznacza typ obiektu przestrzennego. Każdy obiekt przestrzenny z założenia musi posiadać zewnętrzny identyfikator, unikalny przynajmniej w ramach całej infrastruktury INSPIRE (atrybut *inspireId* typu *Identifier*, zdefiniowanego w typach podstawowych INSPIRE). Identyfikator INSPIRE składa się z lokalnego identyfikatora, przestrzeni nazw i identyfikatora wersji. Geometria wywodzi się z hierarchii geometrii zdefiniowanych w języku GML (ISO 19136). Każdy obiekt przestrzenny może mieć przypisanych dowolnie wiele nazw geograficznych. Poza tym obiekty przestrzenne mogą mieć dowolnie dużo innych właściwości (atrybutów i asocjacji).



Rys. 2. Przykład użycia nazw geograficznych
Fig. 2. An example of geographical name usage

Na rys. 3 przedstawiono uproszczony model gazetera INSPIRE [4], będącego modyfikacją modelu z normy ISO 19112. Klasa *LocationInstance* odpowiada w pewnym przybliżeniu klasie *NamedPlace* z rys. 2, w której nazwy geograficzne zostały zredukowane do zapisu tek-

stowego i włączone do klasy głównej. Uprzywilejowanie jednej z nazw powoduje, że, w sytuacjach gdy istnieje wiele oficjalnych nazw danej encji przestrzennej, konieczne jest utworzenie wielu instancji klasy *LocationInstance*.



Rys. 3. Uproszczony model danych INSPIRE dla gazetera [4]
Fig. 3. Simplified INSPIRE data model for gazetteer [4]

Prowadzonych jest wiele prac nad standaryzacją i udostępnianiem nazw geograficznych. Model EuroGeoNames wzbogaca klasę *LocationInstance* z gazetera o atrybuty pozwalające przechować referencję do nagrania wymowy oraz transkrypcję wymowy. W projekcie GeoNames [10] stosowana jest jedna tabela, której kolumny przechowują atrybuty nazw geograficznych. W tym modelu nazwy geograficzne w różnych językach są przechowywane bez informacji o języku, w którym zostały zapisane.

Z porównania modeli wynika, że model danych INSPIRE dla nazw geograficznych pozwala na przechowanie największej ilości danych i charakteryzuje się największym stopniem normalizacji ich struktury. Są w nim wprost przechowane informacje o języku nazwy oraz jej transkrypcji, a w celu przechowywania różnych nazw danej encji nie trzeba wprowadzać sztucznie wielu obiektów, z których każdy ma tę samą geometrię i pozostałe atrybuty z wyłączeniem nazwy.

3. Wymagania względem wyszukiwania po nazwach geograficznych

W odróżnieniu od wielu istniejących rejestrów nazw geograficznych, w INSPIRE zdefiniowano typ danych reprezentujący nazwę geograficzną, który to typ może być wykorzysty-

wany do definicji atrybutów wielu różnych typów danych przestrzennych, np. jednostek administracyjnych, miejsc chronionych, budynków itd. Powoduje to, że nazwy geograficzne są przechowywane w różnych zbiorach danych. Z założenia dane przestrzenne w ramach infrastruktury informacji przestrzennej INSPIRE są rozproszone. W konsekwencji, w przypadku wyszukiwania obiektów przestrzennych na podstawie warunków na nazwach geograficznych warto rozważyć dwa przypadki:

- wszystkie dane zostały zgromadzone w jednej bazie danych,
- dane są dostępne w różnych składnicach danych, dostępnych za pośrednictwem usług sieciowych.

Ważnym aspektem wyszukiwania według nazw geograficznych jest różnorodność wykorzystywanych typów danych:

- dane tekstowe, w tym wielojęzyczne,
- listy kodowe i wyliczenia,
- transkrypcje,
- nagrania wymowy,
- położenie w przestrzeni nazywanego obiektu (jego geometria).

Wyszukiwanie powinno umożliwić zadawanie warunków na każdy z tych typów danych.

Wydaje się, że podstawowym wzorcem dostępu do danych, w przypadku wyszukiwania opartego na nazwach geograficznych, jest mniej lub bardziej dokładna znajomość zapisu lub wymowy nazwy miejsca w jednym z języków. Dodatkowo zadający zapytanie na ogół może zgrubnie zawęzić obszar poszukiwań, zadając prostokąt ograniczający lub podając obszar administracyjnych bądź geograficzny, którego ma dotyczyć zadanie wyszukiwania. W konsekwencji podstawowym rodzajem wyszukiwania jest wyszukiwanie tekstowe, posiadające jednak swoją specyfikę:

- nazwy są stosunkowo krótkimi tekstami,
- część nazw jest wielowyrazowa i mogą być w nich stosowane skróty,
- wiele miejsc może posiadać więcej niż jedną nazwę oficjalną lub może być ona różnie zapisywana [5],
- podczas zadawania zapytań często może dochodzić do przekręcenia nazw lub będą wykorzystywane tylko litery alfabetu angielskiego.

W przypadku predykatów przestrzennych można się spodziewać warunków typu okno, tzn. użytkownik zadaje wprost lub rysuje na mapie prostokąt ograniczający obszar wyszukiwania. Można się również spodziewać ograniczeń przestrzennych przez podanie ogólnie znanego obiektu (np. państwo, kraina geograficzna), na obszarze którego ma leżeć szukany obiekt, lub podanie maksymalnej odległości, w jakiej ma leżeć od zadanego obiektu.

Z założenia dane w ramach infrastruktury INSPIRE są zapisywane na podstawie kodowania UTF-8 i takie kodowanie powinno zostać wykorzystane podczas zadawania zapytania, jego wykonywania oraz prezentacji wyników. Domyślnym sposobem zapisu jest zapis w plikach XML, zgodnie ze schematami XSD z odnośnych specyfikacji INSPIRE.

W ramach infrastruktury informacji przestrzennej, tworzonej w ramach INSPIRE, będą rejestrowane nazwy geograficzne obiektów przestrzennych należących do tematów wymienionych w aneksach Dyrektywy INSPIRE. Liczba takich obiektów i ich nazw jest trudna do oszacowania. Przykładowo w Polsce prowadzony jest *Państwowy rejestr nazw geograficznych* [11], zawierający ok. 100 tys. nazwanych obiektów. Z kolei rejestr Poland Topographic Map zawiera ok. 45 tys. nazw [12].

Największym rejestrem, jaki udało się znaleźć, jest GeoNames, zawierający ok. 7,5 mln. nazwanych obiektów przestrzennych i ok. 10 mln. nazw [10]. Rejestr ten integruje nazwy z różnych rejestrów krajowych, organizacji itd. Najwięcej nazw pochodzi z USA – ok. 2 mln., gdzie przypada ok. 7 nazw na 1000 mieszkańców. W Europie Zachodniej jest to ok. 5 nazw na 1000 mieszkańców. Przyjmując takie proporcje, można się spodziewać, że w Unii Europejskiej, zamieszkałej przez ok. 500 mln. mieszkańców, zostaną zgromadzone nazwy ok. 2,5 mln. obiektów przestrzennych, z których każdy może posiadać nazwy w wielu językach. Dodatkowo mogą być rejestrowane nazwy historyczne. Może to dać docelowo liczbę nazw rzędu kilkunastu lub nawet kilkudziesięciu milionów.

Ze względu na ogólność modelu danych nazw geograficznych INSPIRE, można się spodziewać, że zostanie on lub jego wersja przyjęta oficjalnie (standard de facto) lub zaakceptowana (standard de jur) przez społeczność zajmującą się nazwami geograficznymi. W konsekwencji może być przechowywanych znacząco więcej nazw niż wynikałoby to z zakresu Dyrektywy INSPIRE.

4. Funkcjonalność baz danych w kontekście nazw geograficznych

W niniejszym rozdziale przedstawiono możliwości wyszukiwania oferowane przez wiodące systemy zarządzania bazami danych (SZBD) w zakresie wymaganym przez obsługę nazw geograficznych przechowywanych zgodnie z modelem danych INSPIRE. W zakresie komercyjnych baz danych przedstawiono produkty Oracle, DB2 i SQL Server. W zakresie rozwiązań otwartego oprogramowania wybrano MySQL i PostgreSQL, wraz z towarzyszącym mu rozszerzeniem przestrzennym PostGIS. SZBD Oracle omówiono w wersjach 11g Release 2, SZBD DB2 w wersji 9.7, SZBD SQL Server w wersji 2008 [8], PostgreSQL w wersji 9.0 [21], wraz z rozszerzeniem PostGIS w wersji 1.5.2 [20], a MySQL w wersji 5.5 [16].

Opis dostępnych funkcji oparto na dokumentacji systemów oraz wiedzy autorów. W zestawieniach dostępność danej funkcji oznaczono symbolem +, brak dostępności symbolem –, a okrojoną funkcjonalność symbolem +/–.

4.1. Wyszukiwanie tekstowe i XML

Przedstawione w poprzednim rozdziale wymagania na wyszukiwanie nazw geograficznych w dużej mierze eliminują możliwość oparcia wyszukiwania na operatorze *Like* z SQL i powodują konieczność wykorzystania bardziej zaawansowanych rozwiązań, dostępnych w ramach rozszerzeń do przetwarzania tekstów w SZBD.

Interesujące, dodatkowe elementy wyszukiwania to:

- wyszukiwanie przybliżone (ang. *fuzzy*) – tekstowe wyszukiwanie przybliżone, w którym zwracane są wyrazy podobnie brzmiące,
- transkrypcje – zapis fonetyczny,
- wsparcie dla gromadzenia tych samych informacji w różnych językach w ramach jednego obiektu,
- indeksowanie danych tekstowych celem przyspieszenia wyszukiwania,
- wparcie dla przechowywania danych XML,
- wparcie w SQL dla SQL/XML, XQuery i XPath.

Tabela 1

Porównanie możliwości SZBD w zakresie wyszukiwania tekstowego i XML w kontekście nazw geograficznych

Kryterium	**Oracle**	**DB2 Net Search Extender**	**SQL Server**	**PostgreSQL**	**MySQL**
Wyszukiwanie tekstowe					
Operator *contains* lub analogiczny	+	+	+	+	+
Wyszukiwanie przybliżone	+	+	–	–	–
Transkrypcje	+	+	+/–	–	–
Wsparcie dla danych wielojęzycznych	–	–	–	–	–
Indeksy tekstowe	+	+	+	+	+
XML					
Typ XML	+	+	+	+	–
SQL/XML	+	+	+ (MS SQLXML)	+	–
XPath	+	+	+	+	+
XQuery	+	+	+	–	–

Wszystkie edycje SZBD Oracle udostępniają moduł Oracle Text i podstawową obsługę XML [19]. Dotyczy to również darmowej wersji XE, która jednak nie jest rozwijana i w konsekwencji nie są do niej dodawane nowe funkcje w zakresie obsługi danych tekstowych i XML. W szczególności XE nie udostępnia XQuery. Firma IBM oferuje dwa rozwiązania w zakresie wyszukiwania tekstowego DB2 Net Search Extender [14] i DB2 Text Search [15]. Ze względu na profil wymaganej funkcjonalności, w zestawieniu uwzględniono DB2 Net Search Extender. W tabeli 1 zestawiono możliwości SZBD w zakresie wyszukiwania tekstowego i XML istotne w kontekście nazw geograficznych.

Systemy umożliwiające wyszukiwanie przybliżone i wykonywanie transkrypcji często są ograniczone do najbardziej rozpowszechnionych języków i wykorzystanie ich do innych języków wymaga opracowania dodatkowych słowników.

Ciekawym mechanizmem oferowanym przez SZBD Oracle jest możliwość rejestracji w bazie danych URL-i dokumentów i ich tekstowe indeksowanie.

4.2. Wyszukiwanie przestrzenne

W bazach danych przestrzennych rozróżnia się dane przestrzenne i dane geodezyjne (ang. *geodetic data*), nazywane również danymi geograficznymi (ang. *geographic data*). Zapis i operacje na pierwszym rodzaju danych nie uwzględniają krzywizny kuli ziemskiej w przeciwieństwie do zapisu i operacji na danych geodezyjnych, które są zapisywane i przetwarzane w układach współrzędnych geodezyjnych, np. WGS84. Rozróżnienie to jest istotne z punktu widzenia przetwarzania nazw geograficznych, ponieważ współrzędne wielu obiektów przestrzennych w ramach infrastruktury INSPIRE są zapisywane w układach współrzędnych geodezyjnych. W przypadku zapytań dotyczących nazw geograficznych, duże znaczenie ma możliwość konwersji układów współrzędnych ze względu na heterogeniczność danych.

Firma Oracle oferuje obsługę danych przestrzennych w dwóch wersjach [18]. Oracle Locator jest dostępny we wszystkich edycjach SZBD, ale jego możliwości są ograniczone do przechowywania danych geometrycznych i wykonywania na nich podstawowych zapytań. Oracle Spatial, będący płatnym dodatkiem do edycji Enterprise, rozszerza Oracle Locator o rozbudowane analizy przestrzenne, przetwarzanie danych topologicznych, analizy sieci etc. Firma IBM w ramach SZBD DB2 oferuje obsługę danych przestrzennych w ramach modułu Spatial Extender, dostępnego standardowo we wszystkich edycjach SZBD [13]. Moduł ten pozwala na przetwarzanie danych przestrzennych bez uwzględniania krzywizny Ziemi. Takie funkcje oferuje moduł Geodetic Extender, który jest płatnym dodatkiem do edycji Enterprise. SQL Server 2008 udostępnia przetwarzanie danych przestrzennych i geodezyjnych we wszystkich edycjach. W przypadku bazy danych PostgreSQL, przetwarzanie danych przestrzennych jest obsługiwane przez rozszerzenie PostGIS.

W tabeli 2 zestawiono właściwości baz danych przestrzennych, istotne z punktu widzenia wyszukiwania, z warunkami przestrzennymi, istotnymi w kontekście nazw geograficznych. W niektórych SZBD dla danych geodezyjnych można badać zachodzenie tylko niektórych relacji topologicznych.

Tabela 2

Porównanie możliwości SZBD w zakresie wyszukiwania przestrzennego w kontekście nazw geograficznych

Kryterium	**Oracle Locator**	**Oracle Spatial**	**DB2 Spatial Extender**	**DB2 Geodetic Extender**	**SQL Server**	**PostgreSQL z PostGIS**	**MySQL**
Przetwarzanie danych geometrycznych							
Przechowywanie danych geometrycznych	+	+	+	+	+	+	+
Relacje topologiczne	+	+	+	+	+	+	+
Odległość	+	+	+	+	+	+	+
Najbliżsi sąsiedzi	+	+	–	–	–	–	–
Przetwarzanie danych geodezyjnych (geograficznych)							
Przechowywanie danych geodezyjnych	+	+	–	+	+	+	–
Relacje topologiczne	+	+	–	+	+	+/–	–
Odległość	+	+	–	+	+	–	–
Najbliżsi sąsiedzi	+	+	–	–	–	–	–
Ręczna konwersja układów współrzędnych	–	+	–	+	–	+	–
Automatyczna konwersja układów współrzędnych	+	+	–	–	–	–	–

Wszystkie serwery oferują indeksy przestrzenne, więc to kryterium nie zostało włączone do zestawienia. PostGIS dla danych geodezyjnych udostępnia relacje topologiczne *Intersects*, *Covers* i *CoveredBy*, co w przypadku nazw geograficznych wydaje się wystarczające w typowych zapytaniach. Oracle Locator oferuje wiele wyspecjalizowanych operatorów, takich jak wyszukiwanie najbliższych sąsiadów, którzy znajdują się nie dalej niż zadana odle-

głość, oraz znajdowanie wszystkich obiektów w zadanej odległości. Zapytania takie można wyrazić przy użyciu warunków na odległość, jednak wykorzystanie dedykowanych operatorów może poprawić wydajność przetwarzania zapytań.

4.3. Wyszukiwanie audio

Z tabeli 3 wynika, że moduły SZBD, obsługujące nagrania dźwiękowe na poziomie sygnałów audio, należą do rzadkości. Jeżeli obsługa dźwięków jest dostępna, to sprowadza się do stworzenia biblioteki multimediów z tytułami, wykonawcami itp. przypisanymi do plików dźwiękowych. Atrybuty te służą następnie do filtrowania danych. Rozszerzenia dźwiękowe dostosowane są więc to do wyszukiwania muzyki bez uwzględnienia sygnału audio. Ekstrakcja cech ogranicza się do atrybutów plików dotyczących formatu audio (np. [17]).

Tabela 3

Porównanie możliwości SZBD w zakresie wyszukiwania dźwięków w kontekście nazw geograficznych

Kryterium	Oracle Multimedia	DB2 Audio Extender	SQL Server	PostgreSQL	MySQL + Audio DB [9]
Ekstrakcja cech	+/– (metadane pliku zapisywane do specjalnego obiektu)	– (głównie informacje o pliku)	–	–	–
Wyszukiwanie po dźwiękach	– (po metadanych)	– (po powyższych atrybutach)	–	–	–

Z powyższego wynika wniosek, że aby wyszukiwanie w zbiorze dźwięków było dostępne w bazie danych, należy użyć dodatkowego, zaawansowanego narzędzia do analizy sygnałów. Od bazy danych wymagane byłoby przechowanie plików dźwiękowych (jako danych binarnych) lub jedynie adresów plików na serwerze lokalnym lub w sieci. W tym miejscu wymagane jest podjęcie wszelkich starań, by dostęp do plików był jak najszybszy i by podczas pobierania próbek wykorzystać wszelkie dodatkowe warunki zapytania, ograniczające ich liczbę.

By umożliwić wyszukiwanie po wymowie nazw geograficznych, dobrym pomysłem będzie dostosowanie istniejącego algorytmu do potrzeb systemu. Interesujące nagrania będą krótkie, a wyszukiwanie będzie następowało po całej zawartości, a nie po fragmencie jak w typowych bazach multimedialnych. W pracach [1, 6, 7] zaproponowano różne podejścia do ekstrakcji cech sygnałów dźwiękowych oraz wykorzystania ich w zapytaniach „dźwiękowych". Metadane nagrań dźwiękowych mogą być oparte na widmach częstotliwościowych sygnału, mocy sygnału, współczynniki MFCC (ang. *Mel-Frequency Cepstral Coefficients*) [1] itp. Możliwe jest zastosowanie kompresji do określania podobieństwa [6] – sygnały podobne skompresują się lepiej razem niż osobno. W [7] autorzy proponują utworzenie modelu funkcji

gęstości prawdopodobieństwa dla krótkich fragmentów nagrań i określenie ich podobieństwa z wykorzystaniem klasyfikatorów. Jest to dobry punkt wyjścia do implementacji wyszukiwania nazw geograficznych. Dodatkowym ułatwieniem jest fakt, że fragmenty, na których operują powyższe algorytmy, są krótkie. Nazwy geograficzne będą raczej bardzo krótkimi nagraniami – w jednym nagraniu lektor wypowie maksymalnie kilka słów. Jeżeli jednak nagrania wzorcowe zostałyby utworzone przez lingwistów, to pojawia się pytanie, czy dykcja przeciętnego użytkownika wystarczy do poprawnego dopasowania wypowiedzianej przez niego nazwy do nagrań wzorcowych? Problemem może być również wpływ takich wzorcowych nagrań na ekstrakcję cech.

5. Optymalizacja zapytań z warunkami dotyczącymi nazw geograficznych

Rozważając projekt bazy danych przechowującej dane przestrzenne INSPIRE, należy rozważyć wiele aspektów. Z punktu widzenia zadań wyszukiwania opartych na predykatach dotyczących nazw geograficznych, ważne są następujące zagadnienia:

- nazwy geograficzne są atrybutami złożonymi, zapisywanymi domyślnie w XML i mogą być przechowywane w różny sposób, w bazach danych oferujących przechowywanie i przetwarzanie danych XML i przestrzennych,
- nazwy geograficzne są atrybutami różnych typów obiektów przestrzennych.

Jeżeli chodzi o pierwsze zagadnienie, to można wskazać kilka rozwiązań:

- obiekty przestrzenne są przechowywane w postaci dokumentów XML,
- atrybuty obiektów przestrzennych są przechowywane jako wartości kolumn znormalizowanego modelu relacyjnego, z wykorzystaniem typów przestrzennych,
- kombinacja powyższych podejść, w której dane są dekomponowane i częściowo przechowywane w postaci typów prostych, a częściowo w postaci dokumentów XML.

Jeżeli chodzi o drugie zagadnienie, to można wyróżnić następujące podejścia:

- nazwy geograficzne każdego z typów przestrzennych są przechowywane w dedykowanym komplecie tabel,
- wszystkie nazwy geograficzne są przechowywane w jednym komplecie tabel i specjalna kolumna dyskryminatora określa, w której tabeli są przechowywane opisywane obiekty.

Z punktu widzenia efektywności przetwarzania, wskazana wydaje się dekompozycja, ponieważ umożliwia ona pełniejsze wykorzystanie indeksów oraz operacji udostępnianych przez SZBD. Przykładowo, zapis wprost geometrii obiektów umożliwia założenie na nich indeksu przestrzennego oraz wykorzystanie w zapytaniach SQL predykatów przestrzennych.

Z kolei zapis wartości atrybutów o dziedzinach określonych przez listy kodowe w dedykowanych kolumnach umożliwi zwiększenie selektywności indeksów w stosunku do ogólnych indeksów tekstowych lub indeksów dla typów XML. Podobnie założenie indeksów tekstowych tylko na kolumnach przechowujących pisownię nazw geograficznych powinno poprawić ich selektywność.

Ułatwienie zapisu zapytań z warunkami tylko na nazwy geograficzne oraz zapewnienie największej wydajności ich przetwarzania zapewniłoby zgromadzenie wszystkich danych w jednej bazie i zapis wszystkich nazw geograficznych w jednym komplecie tabel implementujących model danych INSPIRE dla nazw geograficznych. Podejście takie jest analogiczne do założeń gazetera, jednak implementacja modelu INSPIRE dla nazw geograficznych umożliwia przechowywanie i przeszukiwanie znacząco większej ilości informacji.

Zapis i optymalizację zapytań komplikuje fakt, że dane przestrzenne mogą być zapisane w różnych układach odniesienia. SZBD wymagają na ogół, aby dane zapisane w jednej kolumnie były w tym samym układzie odniesienia. W konsekwencji obiekty jednego typu przestrzennego mogą być przechowywane w różnych tabelach.

Standardowe SZBD nie wspierają wyszukiwania po dźwiękach. Należy je więc zaimplementować jako ich rozszerzania lub w aplikacji. Porównywanie sygnałów dźwiękowych może być czasochłonne. Oprócz nagrań należałoby przechowywać ich charakterystykę (metadane sygnałowe). Umożliwiłoby to dwuetapową selekcję na podstawie predykatów na nagraniach. W pierwszym etapie byłyby ewaluowane warunki na metadanych sygnałowych oraz pozostałe warunki z zapytania, których ewaluacja charakteryzuje się niskim kosztem, co pozwoliłoby wyłonić zbiór obiektów, które potencjalnie spełniają warunki zapytania. W drugim etapie, dla obiektów z tego zbioru byłyby sprawdzane pozostałe, bardziej kosztowne warunki. Kolejność tych warunków byłaby ustalana przez optymalizator na podstawie oszacowania kosztów i selektywności.

W modelu danych INSPIRE dla nazw geograficznych wymowa jest przechowywana jako identyfikatory zasobów URI. W celu zwiększenia efektywności przetwarzania, korzystne wydaje się zapisanie nagrań wraz z danymi w bazie danych. Jeżeli nie byłoby to możliwe, np. ze względu na ograniczenia prawne, wtedy koszty dostępu do nagrań musiałby być uwzględnione przez optymalizator bazy danych. W takim przypadku jeszcze ważniejsze byłyby metadane sygnałowe.

Jako ogólną regułę heurystyczną optymalizacji zapytań z warunkami na nazwy geograficzne proponuje się przyjąć zapis jak największej części zadania wyszukiwania w postaci pojedynczego zapytania SQL i pozostawienie optymalizatorowi kosztowemu możliwości wyboru najlepszej strategii jego wykonania. W tym celu ważne są rozszerzenia SZBD w zakresie przetwarzania danych przestrzennych, tekstowych, dźwiękowych i XML. W przypadku

ich braku, konieczne jest napisanie brakującego kodu, który należy dołączyć do SZBD lub, jeżeli nie jest to możliwe, umieścić w aplikacji i wykonywać jako dodatkowy etap realizacji zapytania.

6. Posumowanie

Nazwy geograficzne są jednym z podstawowych typów danych referencyjnych w infrastrukturze informacji przestrzennej budowanej w ramach INSPIRE i są wykorzystywane w modelach danych dla wielu tematów z aneksów Dyrektywy INSPIRE [2]. Ze względu na dużą liczbę nazw geograficznych (rzędu milionów, a nawet dziesiątek milionów) i złożoną strukturę ich opisu, efektywność wyszukiwania danych przestrzennych po nazwach geograficznych jest istotna.

W pierwszej części niniejszego artykułu przedstawiono model danych dla nazw geograficznych, opracowany w ramach INSPIRE [5], i zestawiono go z modelami wykorzystywanymi w gazeterach i geoportalach poświęconych nazwom geograficznym. Zdaniem autorów model INSPIRE jest najpełniejszym modelem danych nazw geograficznych i można się spodziewać, że zostanie wykorzystany do harmonizacji danych przestrzennych również poza infrastrukturą INSPIRE.

W drugiej części przedstawiono wymagania względem wyszukiwania po nazwach geograficznych. W trzeciej części przedstawiono możliwości realizacji tych wymagań przez wiodące komercyjne i dostępne na zasadach otwartego oprogramowania systemy zarządzania bazami danych (SZBD). Z wykonanego przeglądu SZBD wynika, że żaden z nich nie spełnia wszystkich zidentyfikowanych wymagań w zakresie wyszukiwania po nazwach geograficznych, głównie ze względu na dane dźwiękowe. W pozostałym zakresie wymagania są w większości spełnione przez komercyjne rozwiązania w edycjach *Enterprise* z płatnymi rozszerzeniami, co jest kosztowne. W przeglądzie najgorzej wypadł MySQL, któremu brakuje wsparcia dla danych geodezyjnych i dokumentów XML. Z przeprowadzonych prac wynika, że najlepsze wsparcie dla danych przestrzennych i tekstowych w kontekście nazw geograficznych oferuje Oracle. Posiadanie najdroższych, komercyjnych wersji SZBD nie zapewnia jednak poprawnej obsługi danych wielojęzycznych – dostarczana obsługa dotyczy głównie języków: angielskiego, francuskiego i niemieckiego. Implementując wyszukiwanie po nazwach geograficznych, należy liczyć się z koniecznością implementacji części wyszukiwania w aplikacji. W rozdziale piątym zaproponowano reguły podziału przetwarzania pomiędzy aplikację i SZBD oraz projektowania schematu bazy danych, przechowującej nazwy geograficzne.

Dalsze prace będą dotyczyć eksperymentalnej oceny efektywności zaproponowanych rozwiązań na podstawie przedstawionych kryteriów.

Podziękowania

Praca była współfinansowana ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

BIBLIOGRAFIA

1. Cont A., Dubnov S., Assayag G.: Guidage: A Fast Audio Query Guided Assemblage. Proceedings of the International Computer Music Conference (ICMC), 2007.
2. DYREKTYWA 2007/2/WE PARLAMENTU EUROPEJSKIEGO I RADY z dnia 14 marca 2007 r., ustanawiająca infrastrukturę informacji przestrzennej we Wspólnocie Europejskiej (INSPIRE).
3. ROZPORZĄDZENIE KOMISJI (UE) nr 1089/2010 z dnia 23 listopada 2010 r. w sprawie wykonania dyrektywy 2007/2/WE Parlamentu Europejskiego i Rady w zakresie interoperacyjności zbiorów i usług danych przestrzennych.
4. INSPIRE Generic Conceptual Model. Wersja 3.3 z 2010-06-18.
5. D2.8.I.3 INSPIRE Data Specification on Geographical Names – Guidelines. Wersja 3.0.1 z 2010-04-26.
6. Helén M., Virtanen T.: A Similarity Measure for Audio Query by Example Based on Perceptual Coding and Compression. 10th International Conference on Digital Audio Effects (DAFx-07), September 2007.
7. Helén M., Virtanen T.: Audio Query by Example Using Similarity Measures between Probability Density Functions of Features. EURASIP Journal on Audio, Speech, and Music Processing, 2010, p.1-1, January 2010.
8. http://msdn.microsoft.com/en-us/library/bb545450.aspx (dostęp 2011-01-26).
9. http://search.cpan.org/~twh/Audio-DB-0.01/ (dostęp 2011-01-26).
10. http://www.geonames.org/ (dostęp 2011-01-26).
11. http://www.codgik.gov.pl/index.php?option=com_content&view=article&id=162&Itemid=116 (dostęp 2011-01-26).
12. http://www.igipz.pan.pl/ksig/digital_map/home.htm (dostęp 2011-01-26).
13. IBM DB2 9.7 for Linux, UNIX, and Windows Spatial Extender and Geodetic Data Management Feature User's Guide and Reference Updated September, Version 9, Release 7, 2010.
14. IBM DB2 9.7 for Linux, UNIX, and Windows Net Search Extender Administration and User's Guide Updated September, Version 9, Release 7, 2010.

15. IBM DB2 9.7 for Linux, UNIX, and Windows DB2 Text Search Guide Updated September, Version 9, Release 7, 2010.
16. MySQL 5.5 Reference Manual. Oracle, 2011.
17. Oracle Multimedia: Managing Multimedia Content. An Oracle White Paper, September 2009.
18. Oracle® Spatial Developer's Guide 11g Release 2, E11830-07, October 2010.
19. Oracle® Text Application Developer's Guide 11g Release 2, E16594-01, August 2010.
20. PostGIS 1.5.2 Manual.
21. PostgreSQL 9.0.3 Documentation. The PostgreSQL Global Development Group, 2010.

Recenzent: Prof. dr hab. inż. Mieczysław Muraszkiewicz

Wpłynęło do Redakcji 31 stycznia 2011 r.

Abstract

The directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 established an Infrastructure for Spatial Information in the European Community (INSPIRE) [2]. One of the most fundamental data in INSPIRE infrastructure are geographical names (GN) extensively used in every-day communication and considered as reference data for other themes connected with environment, area management, and many others [5]. In regard of possible extensive number of GN (millions or tens of millions) and their complex internal structure, the efficiency of searching of spatial data based on GN is vital.

The first part of this paper presents the INSPIRE data model for GN and compares it with models used in gazetteers and geoportals dedicated to GN. The INSPIRE data specification seems to incorporate the most sophisticated GN data model, which is likely to be used outside of the INSPIRE scope.

The second part of the paper deals with the requirements regarding searching by GN. The third part shows the capabilities of fulfilling these requirements by leading commercial and open source database management systems (DBMS). The review showed that none of the evaluated DBMS-es supports all of the identified criteria, mainly because of the inability to include predicates on sounds in the search. Other requirements are usually satisfied by the commercial solutions in enterprise editions with expensive extensions. MySQL turned out to be the worst prepared database to support GN, because of lack of support for geodetic information and XML documents. Oracle, on the other hand, offers the best support for spatial and

text data in the context of GN. Nevertheless obtaining even the most pricy commercial version of a DBMS does not guarantee correct handling of multilingual data. Typical solutions support mainly English, French and German languages. Implementation of GN based search will require to process part of the search in the application. In the chapter 5 the rules of the division of the processing between an application and a DBMS and database schema design considerations are presented.

Further work will consider experimental efficiency evaluation of the proposed solutions based on the presented criteria.

Adresy

Piotr BAJERSKI: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, piotr.bajerski@polsl.pl.
Tomasz PŁUCIENNIK Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, tomek.pluciennik@gmail.com.