POLSKIE TOWARZYSTWO INFORMACJI PRZESTRZENNEJ
ROCZNIKI GEOMATYKI 2013 ❍ Tom XI ❍ Zeszyt 5(62)

PROBLEMY PROJEKTOWANIA I IMPLEMENTACJI SYSTEMÓW DLA ARCHIWÓW DOKUMENTÓW INDEKSOWANYCH GEOGRAFICZNIE

THE DESIGN AND IMPLEMENTATION PROBLEMS WITH SYSTEMS FOR ARCHIVES OF GEOGRAPHICALLY INDEXED DOCUMENTS

Roland Rusiecki

Politechnika Gdańska, Wydział Elektroniki, Informatyki i Telekomunikacji
Katedra Systemów Geoinformatycznych

Słowa kluczowe: archiwa dokumentów, indeksowanie geograficzne, zakres geograficzny
Keywords: archives documents, geographic indexing, geographic scope

Wstęp

W tradycyjnych bibliotekach lub archiwach dokumentów nadawanie indeksów, umożliwiających wyszukiwanie, opiera się zazwyczaj na temacie, autorze, tytule oraz typie dokumentu, jednakże z punktu widzenia wielu dziedzin, równie pożądane wydaje się wyszukiwanie dokumentów za pomocą położenia geograficznego. Autorzy tacy jak: Byron (1987), Hill (1990) oraz RLG (1989) zwrócili uwagę w swoich pracach na konieczność rozwijania i implementacji systemów geograficznego indeksowania dokumentów. Dodatkowo, rozwój sieci informatycznych w ostatnim dziesięcioleciu, stwarza nowe możliwości korzystania z archiwalnych zasobów dokumentów, które przez lata gromadzone były na tradycyjnych nośnikach (papier, kalki dla map, plansze aluminiowe dla pierworysów). Dokumenty mogą zostać poddane cyfryzacji oraz udostępnione w sieci dla upoważnionych odbiorców. Przykładem tego typu zbiorów mogą być archiwa należące do PZGiK (Państwowego Zasobu Geodezyjnego i Kartograficznego) lub archiwa projektów budowlanych. Niniejszy artykuł zawiera analizę funkcjonalną i niefunkcjonalną systemu informatycznego, przeznaczonego do zarządzania dokumentami zorientowanymi geograficznie, oraz przedstawia problemy wynikłe w trakcie prac nad jego realizacją i propozycje ich rozwiązania.

Rozwiązania dotyczące dokumentów indeksowanych geograficznie

Pierwotnie, próby indeksowania geograficznego dokumentów w zbiorach, opierały się na dodatkowym opisie katalogu nazwą lokalizacji geograficznej. Najbardziej znanym tego przykładem jest indeks dokumentów przyjęty w Bibliotece Kongresu Stanów Zjednoczonych – Library of Congress Subject Headings (LCSH). Zakłada on podział dokumentów zgodnie z lokalizacją nadaną w nagłówku np. ART- PARIS, US – HISTORY (Brinker, 1962). Rozwinięciem powyższej strategii indeksowania są systemy udostępniania dokumentów, które automatycznie wyszukują w ich treści nazwy geograficzne i za pomocą tej metody budują indeks geograficzny (Salton, 1989). Jednakże w obu przypadkach, podstawą jest użycie nazwy tekstowej, reprezentującej położenie. Rozwiązanie takie posiada wiele mankamentów, między innymi: niejednoznaczność nazw, zmienność granic geograficznych w czasie, problemy ze stosowaniem neologizmów w nazwach, jak również problemy z różnorodnością w wymowie i pisowni nazw (Griffiths, 1989).

Innym rozwiązaniem problemu indeksowania geograficznego dokumentów, jest zastosowanie współrzędnych geograficznych, do określenia punktu lub obszaru zainteresowania dla dokumentu. Systemy stosujące to rozwiązanie, zorientowane są zazwyczaj na wyszukiwanie informacji przez określanie zakresu na mapie. Przykładem takiego systemu jest Legal Atlas (www.leibnizcenter.org/general/legal-atlas). W przypadku systemów opartych na współrzędnych, problemy stwarza automatyczne określanie zakresu dokumentu na podstawie analizy jego treści (Woodruff, 1994).

Założenia opisywanego systemu informatycznego

W trakcie prac związanych z tworzeniem systemu informatycznego, dedykowanego dla archiwów zorientowanych geograficznie, należy rozważyć poniższe problemy.

1. Do budowy indeksu geograficznego, należy posłużyć się zakresem opartym na współrzędnych geograficznych. Przyjęcie innego rozwiązania, na przykład oparcie georeferencji dokumentu na numerze działki, rodzić może analogiczne problemy, jak budowanie indeksu geograficznego w oparciu o nazwę opisującą położenie (zmienność zakresu w czasie, niejednoznaczność nazw itd.)
2. Dokumenty w archiwum mogą zawierać informacje niejawne, bądź dane osobowe. System informatyczny musi zapewniać kontrolę dostępu, biorąc pod uwagę te dwa kryteria.
3. Każde tradycyjne archiwum dokumentów posiada zazwyczaj własny system informatyczny, informacje w nim zawarte posłużyć mogą do wyszukiwania dokumentu w nowym archiwum elektronicznym. Należy wziąć pod uwagę, czy system taki jest już nieaktywny, zawiera jedynie informacje historyczne, czy jest cały czas „żywym" systemem i informacje w nim się znajdujące należy konsolidować w archiwum elektronicznym.
4. Podział dokumentu w nowym archiwum elektronicznym musi umożliwiać przesyłanie go drogą elektroniczną. Jednocześnie nazewnictwo asortymentu, z którego składa się dokument, musi być jednoznaczne i intuicyjne dla użytkownika końcowego.

5. Mapa, na której wizualizowane są zakresy dokumentów powinna być obrazem, do którego przyzwyczajeni są użytkownicy końcowi. Ułatwia to szybkie orientowanie się i wyszukiwanie dokumentów za pomocą wskazań. Nie wyklucza to stosowania jako dodatkowych warstw jakichkolwiek innych map, dostępnych za pomocą usług serwerów WMS (Web Map Service).

Archiwami, które z definicji zawierają dokumenty związane z lokalizacja przestrzenną, są elementy PZGiK. W dalszej części artykułu nawiązuje się głównie do tego typu zbiorów, jako reprezentatywnych dla kompleksowego omówienia problemu.

Charakterystyka oprogramowania

Oprogramowanie do rozwiązania powyższego problemu, zrealizowane zostało w oparciu o pakiet **JustMap** (autorstwa własnego) współpracujący z otwartą bazą danych **Firebird 2.1.**

Pakiet składa się z trzech części aplikacji:

1. **JustMapEditor** umożliwia wizualną edycję mapy, na zasadach przypominających pracę z narzędziami z rodziny *CAD*. Przyjęcie takiego rozwiązania związane jest z subiektywną opinią autora, że rzesze inżynierów w naszym kraju najbardziej przyzwyczajone są do interfejsu użytkownika jaki posiada **AutoCad**, niż typowy program **GIS**, za jaki poczytywany może być choćby **OpenJump**.

2. Program **JustMapBuilder** służy do zarządzania bazą danych, w tym:
   - tworzenia map z plików w formatach typu GML, SHP, MIF, MID,
   - zarządzania użytkownikami i ich uprawnieniami,
   - tworzenia kopii zapasowych geobazy,
   - zarządzania web serwerem realizowanym przez aplikację internetową **JustMapServer.**

3. Program **JustMapServer** udostępniający aplikację internetową, umożliwiającą przeglądanie mapy za pomocą przeglądarek obsługujących *HTML 5*, co wynika głównie z wykorzystywania znacznika *Canvas*. Oprogramowanie testowane było na przeglądarkach Internet Explorer w wersji powyżej 9, Firefox oraz Gogle Chrome (rys. 1).

Pakiet **JustMap**, w wersji przygotowanej do usługi archiwów, pracuje w oparciu o szerszy katalog obiektów wektorowych niż obiekty przewidziane zaleceniami OGC (Open Geospatial Consortium). Umożliwia to łatwy import plików z formatów typu MIF, DXF posiadających również szerszy katalog obiektów. Niemniej jednak dzięki aplikacji **JustMapBuilder** możliwy jest późniejszy eksport danych do formatów SHP czy GML, zawierających obiekty zalecane przez OGC (Obe, Hsu, 2011).

Założeniem wyjściowym do realizacji oprogramowania, było stworzenie aplikacji przeznaczonej do przeglądania i edycji mapy oraz prowadzenia archiwum dokumentów w formacie PDF, których wyszukiwanie oprzeć można o wskazania na mapie lub klasyczne zapytania wykonywane w oparciu o metadane dokumentów.

Twórcą całości oprogramowania jest autor artykułu, oprogramowanie ma charakter komercyjny, jednak zawarte w publikacji uwagi są na tyle generyczne, że mogą być wzięte pod uwagę w przypadku implementacji podobnych rozwiązań, w jakichkolwiek innych systemach GIS.

Rys. 1. Określanie zakresu przeszukiwań w przeglądarce Google Chrome – JustMapServer

Problemy w trakcie realizacji i sposoby i rozwiązywania

Pozycjonowanie geograficzne dokumentów

Nadawanie pozycji geograficznej dokumentom, odbywa się za pomocą zdefiniowania w geobazie warstwy o nazwie ZAKRDOK. Warstwa za pomocą obiektów typu point, polygon lub linestring (Obe, Hsu, 2011) umożliwia określenie obszaru, którego dotyczy dokument znajdujący się w archiwum.

Relacja pomiędzy dokumentem a warstwą ZAKRDOK, określona została jako jeden do wielu, dzięki czemu dokument (zbiór dokumentów, teczka) może mieć kilka odrębnych zakresów, niebędących geometriami ciągłymi. Możliwość podzielenia zakresu na kilka odrębnych geometrii należy rozumieć jako odzwierciedlenie sytuacji ze świata rzeczywistego, w którym zbiór dokumentów, jakim jest na przykład projekt budowlany, dotyczy kilku odrębnych lokalizacji, np. dwóch odcinków ścieżki rowerowej.

Wyszukiwanie dokumentów, opierając się o położenie geograficzne, polega na wskazaniu w aplikacji pracującej w środowisku przeglądarki internetowej obszaru, którego dotyczyć mają zarchiwizowane dokumenty. Dalszą selekcję dokumentów do przeglądania oprzeć można na nieprzestrzennych metadanych, które można wprowadzić do systemu na etapie cyfryzacji dokumentów papierowych lub pozyskać z istniejących systemów bazodanowych.

Integracja, danych z istniejącymi systemami informatycznymi

Pozyskanie nieprzestrzennych danych opisujących dokument można oprzeć na:

1) wprowadzeniu do systemu na etapie cyfryzacji, na zasadzie analizy samych dokumentów;

2) jednorazowym skopiowaniu danych z istniejących wcześniej systemów informatycznych;

3) integracji geobazy z istniejącym systemem informatycznym.

Pierwsze dwie możliwości dotyczą sytuacji, w której: 1) brak było do tej pory systemu informatycznego zarządzającego archiwum dokumentów lub 2) system informatyczny zakończył swoje funkcjonowanie. W przypadku kiedy istnieje „żywy" system zarządzający archiwum dokumentów, jedynym rozwiązaniem jest integracja obu systemów.

W celu poprawienia czytelności opracowania, w dalszej części baza danych pakietu **JustMap** nazywana będzie **bazą JustMap**, natomiast baza danych jakiegokolwiek istniejącego oprogramowania **bazą zewnętrzną**.

W wariancie integrowania z systemem istniejącym, oprogramowanie **JustMap**: umożliwia wyszukiwanie dokumentów za pomocą zapytań przestrzennych, zapewnia kontrolę poziomów dostępu do dokumentów, udostępnia infrastrukturę do wyszukiwania za pomocą danych nieprzestrzennych, natomiast same dane nieprzestrzenne pobierane są z bazy istniejącego systemu zarządzania archiwum.

Problemem, który należało rozwiązać było ustalenie jednoznacznego klucza dla zbioru dokumentów, jednoznacznego dla obu baz danych. Z uwagi na założenie, że **baza zewnętrzna** jest wykorzystywana jedynie do odczytu danych, w **bazie JustMap** stworzono tabelę przechowującą klucz zbioru dokumentów pochodzący z **bazy zewnętrznej**.

Innym problemem, było przyjęcie unikalnego numeru identyfikującego zbiór dokumentów, który widziany jest przez użytkownika końcowego przeglądającego archiwum lub przez personel wprowadzający dane do systemu. Intuicyjnie podchodząc do problemu, najprostszym rozwiązaniem jest przyjęcie dotychczasowego numeru opisującego teczkę dokumentów w istniejącym archiwum. W zależności od typu archiwum, może to być na przykład: 1) numer pozwolenia na budowę, 2) numer KERG w archiwach należących do PZGiK itd. Z uwagi na fakt, że sposób numeracji może zmienić się w przyszłości, podobnie jak zmieniał się na przestrzeni lat, oprogramowanie umożliwia generowanie numerów dokumentów w oparciu o pola danych, zawarte w bazie zewnętrznej. W przypadku zmiany przepisów. numery te mogą być masowo zmienione w całej bazie, bez ryzyka utraty integralności bazy danych, która opiera się na identyfikatorach GUID niewidocznych dla użytkownika.

Poziomy dostępu do dokumentów

Oprogramowanie, pracując w oparciu o bazę danych FIREBID, ma możliwość dodawania użytkownika na poziomie serwera. Następnie użytkownicy standardowo otrzymują przywileje na poziomie każdej z baz danych.

Z uwagi na fakt, że przechowywane w bazie danych dokumenty mogą posiadać klauzulę poufności i zawierać dane osobowe na poziomie każdej z baz, wprowadzono dodatkowe ograniczenia dostępu:

1) z uwagi na poziom niejawności – *jawne, zastrzeżone, poufne, tajne, ściśle tajne*,

2) z uwagi na występowanie danych osobowych – *zawiera dane osobowe, nie zawiera danych osobowych.*

Wprowadzenie ograniczeń dostępu związanych z klauzulą niejawności, pociągnęło za sobą konieczność zablokowania możliwości dodawania do systemu użytkownika o nazwie takiej samej jak użytkownik wcześniej istniejący a usunięty, w przypadku kiedy użytkownik usunięty posiadał prawa dostępu do informacji niejawnych. Ograniczenie takie wynika ze stosownych przepisów, dotyczących ochrony informacji niejawnych.

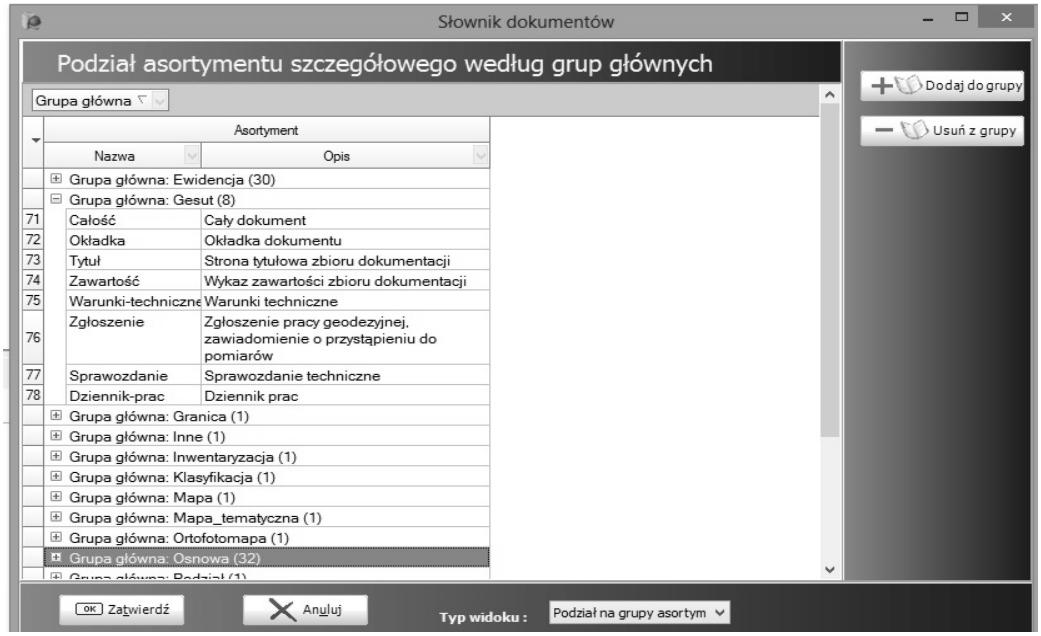
Klauzule niejawności oraz to czy dokument zawiera w sobie dane osobowe, określane jest na etapie cyfryzacji dokumentu. Klauzula może być nadana zbiorczo całemu dokumentowi (np. teczce) lub poszczególnym plikom, wchodzącym w skład dokumentu. Zapobieganie dostępowi do dokumentów osobom nieposiadającym odpowiednich przywilejów, odbywa się na poziomie aplikacji internetowej, udostępniającej archiwa za pomocą przeglądarek.

Słownik hierarchiczny dokumentów i podział dokumentów

Metoda słownikowania dokumentów w pakiecie oprogramowania oparta została na doświadczeniach, wynikających z cyfryzacji dokumentów, należących do PZGiK. Przykładowe słownictwo związane będzie z występującym tam asortymentem.

Przyjęto dwupoziomowy słownik hierarchiczny. Każdy dokument wchodzący w skład zbioru, może przyjąć nazwę pochodzącą z asortymentu szczegółowego. Jednak to, czy dany asortyment szczegółowy jest dostępny dla nadrzędnego typu dokumentu, zależy od przydzielenia uprawnienia do występowania takiego dokumentu w zbiorze.

Hierarchię dokumentów na szczeblu bazy danych zrealizowano za pomocą trzech tabel. Pierwsza z nich zawiera nazwy jakie przyjmować mogą główne zbiory dokumentów, np. „Ewidencja", „Osnowa". Druga tabela zawiera nazewnictwo asortymentu szczegółowego np. „okładka", „spis treści", „szkic osnowy". Trzecia tabela zawiera informacje, czy dany asortyment szczegółowy może wystąpić jako element wybranej pozycji ze zbioru głównego.



Rys. 2. Hierarchiczny słownik dokumentów

Takie rozwiązanie umożliwia elastyczne budowanie słownika, w którym niektóre pozycje, na przykład spis treści, mogą być stosowane w kontekście wszystkich grup głównych asortymentu, a inne jedynie dla specyficznej grupy głównej. Widok interfejsu słownika przedstawia rysunek 2.

Format dokumentów

Formatem przyjętym do przechowywania dokumentów w archiwum jest PDF. Wybór związany jest z możliwością przechowywania dokumentów wielostronicowych. Pierwotnym założeniem jest konieczność przechowywania w jednym pliku dokumentów, które w formie papierowej stanowią nierozerwalną całość, np. ciąg obliczeń. Ponadto format PDF w wersji A (PDF/A) zgodnie ze standardem ISO 32000-1 przeznaczony jest do długotrwałego przechowywania dokumentów (King, 2009).

Z uwagi na konieczność udostępniania dokumentów przez Internet przyjęto następujące ograniczenia rozmiaru plików:

- dokumenty wielostronicowe zawierające tekst mogą posiadać rozmiar nie większy niż 250 KB na stronę tekstu A4, przy zachowaniu czytelności oryginału;
- dokumenty zawierające grafikę, w tym mapy, mogą posiadać rozmiar nie większy niż 500 KB na stronę A4.

W praktyce okazało się, że ograniczenia te pozwalają na wykonanie wiernych kopii dokumentów, które przy wydruku na skutek automatycznych procesów poprawy, takich jak: usuwanie mory itp. potrafią zachować czytelność lepszą od oryginału. W przypadku dokumentów o formacie większym od A4 ograniczenia stosuje się stosownie do iloczynu stron A4 w danym dokumencie, sprowadzając każdy dokument niezależnie od formatu, do tzw. „strony przeliczeniowej A4", która może być również stosowana jako uniwersalny przelicznik wykonanej pracy, pomiędzy wykonawcą dokonującym cyfryzacji archiwów a zamawiającym, zarządzającym archiwum. Aby możliwa była kontrola wykonanych prac i ich zgodności z ograniczeniami po procesie cyfryzacji należy wykonać zestawienie zawierające: 1) liczbę dokumentów, 2) liczbę stron w każdym dokumencie, 3) liczbę stron przeliczeniowych A4, 4) wielkość pliku w KB na stronę przeliczeniową. W niektórych przypadkach, wykonanie takiego zestawienia ręcznie jest niewiele mniej pracochłonne, jak wykonanie samego procesu cyfryzacji. Dlatego dla zarządzających archiwami PZGiK, autor nieodpłatnie udostępnia oprogramowanie do analizy archiwów cyfrowych PDF pod powyższym kątem.

Podsumowanie

W artykule przedstawiono problemy, które w opinii autora należy rozważyć w trakcie organizowania archiwów dokumentów zorientowanych przestrzennie. Propozycje ich rozwiązania są zdaniem autora na tyle generyczne, że można je zaimplementować w jakimkolwiek systemie GIS.

Na zakończenie nadmienić należy, że przedstawione powyżej rozwiązanie informatyczne dedykowane jest wszelkim archiwom zawierającym dokumenty, które można pozycjonować geograficznie. Pomimo kilkukrotnie przytaczanego przykładu PZGiK, intencją autora było, opisanie oprogramowania mogącego służyć jako archiwum dla dokumentów z różnych branż.

Literatura

Brinker B.,1962: Geographic approach to materials in the Library of Congress subject headings. Library Research and Technical Services.

Byron J., 1987: Topographical indexing. The Indexer.

Griffiths A., 1989: SAGIS: A proposal for a Sardinian geographical information system and an assessment of alternative implementation strategies. *Journal of Information Science* vol.15, issue 4-5: 261-267.

Hill L.L., 1990: Access to Geographic Concepts in Online Bibliographic Files: Effectiveness of Current Practices and the Potential of a Graphic Interface. Dissertation. University of Pittsburgh.

King C., 2009: Long live ISO 32000-1. The PDF Standard. *ISO Focus* 4/2009: 24-25. http://www.iso.org/iso/p.24_main_focus.pdf

Obe R.O., Hsu L.S, 2011: PostGIS in Action. Manning Publications CO. USA.

RLG, 1989: Research Libraries Group enters new sphere with georeferencing project. Research Libraries Group News.

Salton G., 1989: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading, MA.

Woodruff A.G., Plaunt C., 1994: GIPSY : Automated Geographic Indexing of Text Documents. *Journal of the American Society for Information Science* 45(9): 645-655.

Abstract

In typical archives, documents are indexed primarily by subject, author, title, and, to a lesser extent, by document type. Adding the possibility of geographic indexing can make the searching process much more cohesive and comprehensive.

This paper describes design and implementation problems with software for managing archives of geographically oriented documents. Apart from information about geographic indexing this paper also describes some issues specific for our country, e.g. work with documents that contain classified information or personal data.

mgr inż. Roland Rusiecki
roland.rusiecki@studiocad.pl