Zastosowania

O przetwarzaniu dużych dokumentów – duże też może być piękne...

Jacek Kmiecik i Marek A. Valenta

Streszczenie

The main task of BPP AGH (Bibliographic List of Staff Publications) application is accumulation, processing and making accessible on-line data concerned publications of all kinds by the staff of AGH Technical University. The project BPP AGH was based on open software: Linux, Apache, PHP, MySQL, TEX. Processing of the database content into the PDF file is performed with ConTEXt.

Od bazy danych, do interaktywnego PDF-a

Zadaniem aplikacji Bibliografia Publikacji Pracowników Akademii Górniczo-Hutniczej (BPP AGH) jest oczywiście oprócz gromadzenia, przede wszystkim przetwarzanie i udostępnianie w systemie *on line* danych o wszelkiego rodzaju publikacjach, których autorami, bądź współautorami, są pracownicy Akademii Górniczo-Hutniczej im. St. Staszica w Krakowie.

Gromadzenie takich informacji leży w gestii Oddziału Informacji Naukowej Biblioteki Głównej AGH, a rezultat tych prac publikowany był dotychczas w postaci roczników (wydawanych drukiem w nakładzie kilkuset sztuk egzemplarzy, w ramach statutowych obowiązków Biblioteki). Przygotowanie do druku tych roczników odbywało się początkowo „ręcznie”, później przy użyciu komputera (głównie na etapie porządkowania i tekstowego wprowadzania informacji), by wreszcie przybrać postać prostych aplikacji bazodanowych. W taki sposób powstało ok. 10 ostatnich Roczników Bibliografii – ostatni opublikowany, drukowany rocznik, opatrzony jest informacją (zasięgiem chronologicznym) „za rok 1998”.

Wcześniejsza wersja bazy, opracowana w Bibliotece Głównej AGH przez pracowników Samodzielnej Sekcji Komputeryzacji BG, oparta była na aplikacji MS Access 97. Aplikacja ta sprawdziła się w lokalnej sieci Biblioteki, umożliwiła sprawną i elastyczną zmianę struktury bazy, modyfikację oraz niezbędną poprawę jej funkcji archiwizujących, a także użytkowych, niestety nie mogła być wykorzystana jako interaktywna baza danych z dostępem poprzez strony WWW. Ograniczenia ekonomiczne jakie narzuca rzeczywistość uczelniana, skłoniły do szukania rozwiązań opartych na programach *public domain* (lub licencji GNU). Oprócz aspektów ekonomicznych, w doborze rozwiązań dużą rolę odgrywały aspekty bezpieczeństwa gromadzonych danych, łatwości ich wprowadzania, obróbki, a także sprawnego udostępniania tych zasobów w sieci Internet.

Przygotowanie aplikacji BPP AGH odbywało się zgodnie z etapowaniem prac przyjętym przy tego typu przedsięwzięciach:

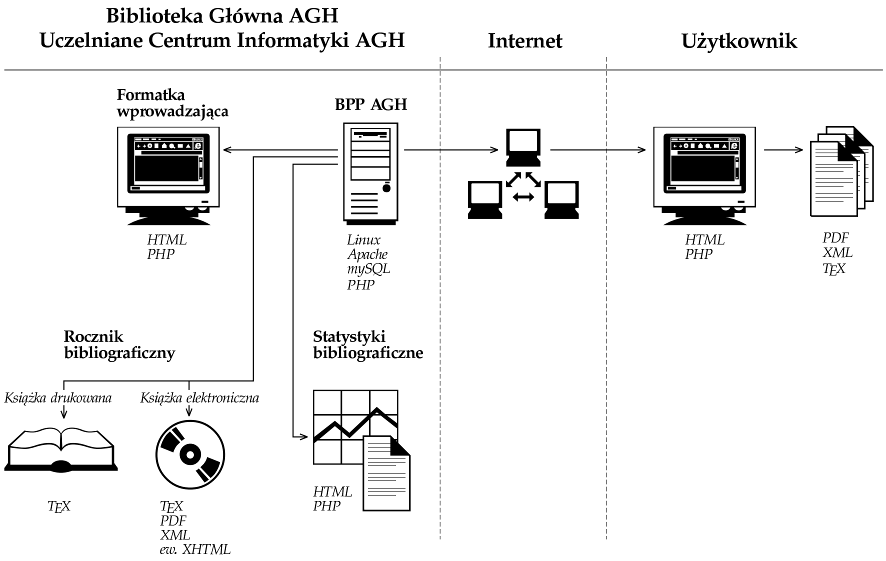
1. opracowanie koncepcji bazy danych i interaktywnej aplikacji do jej przeglądania,
2. dostosowanie struktury bazy do wymogów prezentacji bibliograficznych,
3. projekt i implementacja podsystemu wprowadzania i edycji danych bibliograficznych,
4. projekt i implementacja aplikacji udostępniającej ogólnodostępne informacje bibliograficzne,
5. wykonanie modułu tworzenia statystyk bibliograficznych,
6. opracowywanie *Bibliografii Publikacji Pracowników AGH* w wersji książki elektronicznej.

Ostatecznie system bazy danych eksploatowany jest na jednym z serwerów Uczelnianej Sieci Komputerowej, a dostępny poprzez aplikację WWW dla dwóch typów użytkowników (rys. 1):

1. edytorów danych bibliograficznych,
2. użytkowników końcowych (publicznych).

Edytorami danych jest grupa pracowników Oddziału Informacji Naukowej Biblioteki Głównej AGH. Dla takich typu użytkowników dostęp do funkcji aktualizacji bazy możliwy jest po dokonaniu autoryzacji użytkowników systemu. Autoryzacja użytkownika jest w takim wypadku konieczna celem ochrony bazy przed nieuprawnioną edycją, a w szczególności przez osoby nieposiadające odpowiedniej wiedzy bibliotekarskiej. Mogą one bowiem spowodować dużą ilość błędów merytorycznych w danych, których to błędów mechanizmy integralności bazy danych nie są w stanie kontrolować.

Użytkownikiem końcowym aplikacji BPP AGH może być każdy użytkownik sieci Internet, mający do dyspozycji dowolną przeglądarkę WWW, zgodną ze specyfikacją HTML ver. 4.01 lub nowszą. (Możliwy jest też dostęp poprzez terminale tekstowe i przeglądarki znakowe typu Lynks.)



Rysunek 1: Szkic projektu bazodanowego BPP AGH

System bazy danych oraz całą aplikację opracowano z wykorzystaniem technologii i programów w całości opartych na licencji GNU: Linux, Apache, MySQL, PHP. Powstała architektura wielowarstwowa, w której serwer bazy danych zrealizowano w oparciu o MySQL, serwer aplikacji w oparciu o PHP, a klient wykorzystuje przeglądarkę WWW.

Ze względu na charakter gromadzonych danych i ich przetwarzania dla szerokiego udostępnienia, w realizacji systemu napotkano na wiele problemów, które wymagały szczególnej uwagi i specyficznych rozwiązań. W większości przypadków dotyczyły one elementów edycji złożonych danych tekstowych oraz prezentacji tychże w różnych układach zdeterminowanych przez użytkowników końcowych. Wielokrotnie rozwiązanie tych problemów wiązało się także z koniecznością zmiany formatów przejściowych i końcowych zestawów danych dla użytkownika końcowego (tekst ASCII, TEX, HTML, PDF).

Rekordy bibliograficzne w sporej ilości wpisów zawierają obcojęzyczne znaki diakrytyczne[1], fragmenty formuł matematycznych, symbole chemiczne i fizyczne. Z racji ich występowania oraz kierując się wygodą użytkowników systemu (edytorów, wprowadzających dane do systemu), dla zapisu typów tekstowych w bazie danych zastosowano notację TEX-ową, dającą jednoznaczny zapis w formacie tekstowym (ASCII). Zastosowanie takiego podejścia (tj. formatu TEX) umożliwia dodatkowo, w razie potrzeby, przygotowanie składu Bibliografii bądź jako książki „tradycyjnej” (drukowanej offsetowo), bądź dokumentu elektronicznego (np. hipertekstowy dokument w formacie PDF).

Na etapie tworzenia prezentacji informacji bibliograficznych zastosowano format HTML. Jednak z powodu sporej rozbieżności współczesnych przeglądarek WWW podczas interpretacji tego samego źródła-kodu HTML, wynikła konieczność stosowania ścisłego jego standardu. Założono, że wynik interpretacji kodu winien być jednakowo wyświetlany w co najmniej kilku współcześnie używanych przeglądarkach internetowych.[2]

I tu pojawiło się kilka dodatkowych problemów: format HTML pozwala w bieżącym dokumencie (ściślej, pojedynczej stronie WWW) użyć tylko jednej strony kodowej – stąd wynikły trudności z prawidłowym wyświetlaniem wszystkich obcojęzycznych znaków diakrytycznych. Ponadto, współczesny HTML oferuje wprawdzie możliwość prawidłowego wyświetlania formuł matematycznych, niemniej komplikuje to znacznie kod źródłowy, a jego prezentacja jest możliwa dopiero po zainstalowaniu odpowiednich wtyczek (*plugin*-ów, są to np.: MathML,

[1] Np. à, á, ü, ý, č, š, ç i in. – głównie w tytułach publikacji oraz nazwiskach autorów zagranicznych.

[2] Z ostatnich analiz można wysnuć pozytywne wnioski, iż większość producentów przeglądarek WWW coraz konsekwentniej przestrzega zalecenia standardu HTML – rokuje to optymistycznie dalszym pracom rozwojowym wszelkich aplikacji opatych na tym standardzie.

WebEQ, skrypty Javy). Dopóki nie powstanie jednolity standard prezentacji formuł matematycznych, udokumentowany i zaimplementowany w większości używanych przeglądarek, uważamy, że słusznym jest założenie prezentacji na stronach WWW „surowego" zapisu TEX-wego. Ponieważ dla „niewtajemniczonego" użytkownika systemu zapis taki może wyglądać dość dziwnie, alternatywą obejrzenia tych stron powinna być możliwość ich wyświetlenia jako dokumentu PDF, w którym wyżej wzmiankowane problemy nie będą występować. Taka możliwość prezentacji zawartości stron przewidziana jest do realizacji w trakcie kolejnych etapów rozbudowy aplikacji BPP AGH.

Na podstawie specyfikacji funkcji potrzebnych edytorom danych bibliograficznych oraz kierownictwu Biblioteki Głównej AGH opracowano, a następnie wykonano podsystem wprowadzania i edycji danych oraz moduł umożliwiający dynamiczne przeglądanie statystyk zawartości stworzonej bazy (np. poprzez wybór rocznika, autora oraz jednostki, przy której autor jest afiliowany). Obie aplikacje dostępne są dla dedykowanych grup użytkowników, wyznaczonych przez administratora systemu. Na uwagę zasługuje możliwość wykorzystania modułu sporządzania statystyk, celem weryfikacji merytorycznej bazy i poprawy jej zawartości. W przyszłości moduł ten będzie mógł być rozszerzony o sprawne przeszukiwanie zestawień publikatorskich, np. w celu generowania raportów dla KBN, Komisji Rektorskich, sporządzania sprawozdań wewnętrznych, itp.

Inne problemy trzeba było rozwiązać na etapie opracowywania elektronicznej wersji *Bibliografii Publikacji Pracowników AGH*. Założono bowiem – podpierając się mocnymi argumentami ekonomicznymi – iż począwszy od rocznika „1999", Uczelnia zrezygnuje z wydawania „Bibliografii..." w postaci tradycyjnej książki drukowanej na papierze, tak jak czyniła to dotychczas.

Analiza charakteru danych, wymagań przyszłych użytkowników i dostępnych technologii, wykazała największą przydatność przygotowania wersji elektronicznej tej publikacji w formacie PDF, a dystrybuowanej na nośniku CD lub poprzez Internet.

Format PDF jest najbardziej przenośnym pomiędzy platformami systemowymi i sprzętowymi, pozbawionym większości ograniczeń, jakie niesie za sobą np. format HTML. Posiada kilka niewątpliwych zalet: możliwość definiowania hiperlinków pomiędzy skojarzonymi wzajemnie fragmentami tekstu, możliwość wyświetlania wszystkich znaków diakrytycznych, formuł matematycznych, możliwość „przyzwoitego" drukowania treści dokumentu, i wiele innych.

Powyższy dokument, przy wykorzystaniu systemu BPP AGH powstaje „prawie automatycznie": baza MySQL odpytywana przez skrypty Perlowe (z modułami obsługi baz danych DBI) generuje plik wynikowy w formacie systemu TEX, następnie poprzez dołączenie preambuły z makrami i formatem publikacji (napisany w języku TEX stosowny program formatujący) w wyniku kompilacji programem pdfTEX produkowany jest bezpośrednio dokument PDF. W podobny sposób będzie odbywać się generowanie interaktywne bieżącej strony HTML, oglądanej w przeglądarce WWW, lub dynamicznie wybieranych porcji informacji z bazy danych.

Prace nad systemem BPP AGH trwają – aktualnie uzupełniana jest zawartość bazy o bieżące pozycje bibliograficzne. Z założenia baza zawierać będzie kompletne informacje bibliograficzne od roku 1999. Pozycje dotyczące lat wcześniejszych zamieszczone są sporadycznie, jako suplement do wydanych wcześniej drukiem roczników Bibliografii – dostępne jedynie poprzez Internet.

Kolejnym ważnym etapem prac nad rozwojem bazy będzie poprawa i ujednolicenie zapisu TEX-owego w rekordach bazy – bez tego nie można mówić o jakiejkolwiek automatyzacji generowania dokumentów. O ile przeglądarki WWW dopuszczają występowanie błędów w źródłach HTML-owych, o tyle program TEX nie pozwoli na wystąpienie jakiegokolwiek błędu formalnego w zapisie kodu źródłowego (który generowany jest automatycznie z bazy danych). Z jednej strony jest to dodatkowy aspekt systemowej kontroli poprawności danych, z drugiej jednak staje się niezwykle wymagający dla użytkownika.

Trwają też udane próby zastosowania technologii XML, jako etapu pośredniego pomiędzy pozyskiwaniem danych z bazy, a odpowiednim ich przygotowaniem dla docelowej przeglądarki, lub dla specyficznych sposobów prezentacji tych informacji. XML jest zintegrowany z większością technologii internetowych, a jako format tekstowy stanowi idealną płaszczyznę wymiany danych pomiędzy najprzeróżniejszymi protokołami, niezależnie od systemów operacyjnych czy platform sprzętowych.

W kolejnych etapach rozbudowy aplikacji BPP AGH przewidywane są zatem:

Rocznik	Rozmiar plików w bajtach			Liczba stron	Liczba hyperlinków		Średnia linków*
	TEX	*.tuo	PDF		/Link	/GoTo	
1999	5 838 869	444 074	35 842 706	2761	149 186	125 835	45.5
2000	5 892 629	450 998	36 241 647	2801	150 863	127 446	45.5
2001	6 040 675	460 111	36 962 634	2851	153 698	129 629	45.4

* Średnia liczba linków /GoTo na 1 str. dokumentu PDF

Tabela 1: Zestawienie „wagowe” plików. Objaśnienia: TEX – pliki źródłowe; *.tuo – pliki pomocnicze ConTEXt-a; PDF – pliki wynikowe w formacie PDF

1. kontrola poprawności formalnej kodu TEX-owego podczas redagowania rekordów,
2. sygnalizowane powyżej rozbudowanie funkcji modułu statystyk bibliograficznych,
3. możliwość zmiany sposobu prezentacji danych bibliograficznych wg przyjętych norm i standardów,
4. poprawa i optymalizacja kodu z dostosowaniem do bieżących zmian w technologiach internetowych (XML, XHTML, MathML).

Najcięższa praca? Nie dla TEX-a...

Z TEX-nicznego punktu widzenia, ciekawym etapem prac było przygotowanie wersji elektronicznej PDF trzech roczników Bibliografii, udostępnianej na nośniku CD (lub poprzez sieć Internet).

Wcześniejsza, prototypowa wersja przygotowana została do przetwarzania za pomocą programu pdfTEX. Zrzut z bazy MySQL dokonywany był poprzez skrypty perlowe (z wykorzystaniem modułu DBI). Dzięki zastosowaniu Perla, można było dość swobodnie modelować strukturę pliku wynikowego, który ostatecznie miał być przetwarzany pdfTEX-em. W efekcie końcowym dwa roczniki bibliografii 1999–2000 „ważyły” ok. 40 MB.

Niestety, takie podejście spowodowało, że kolejne wersje pdfTEX-a oraz pakietów makr z nim współpracujących, nie przetwarzały prawidłowo plików przygotowanych według poprzedniej procedury.

Zakładając, że rozwój pdfTEX-a będzie ściśle zintegrowany z coraz bardziej popularnym ConTEXem – drugą wersję elektronicznego wydania BPP przygotowano właśnie pod tym kątem. Skrypty perlowe uległy minimalnej modernizacji, raz: pod względem składni makr ConTEXt-owych, dwa: wydobycia z bazy danych dodatkowych informacji, które docelowo ułatwiły stosowne hyperlinkowanie w obrębie dokumentu.

Ze względu na zwiększenie liczby linków PDF i powiększenie objętości dokumentu[3] – zdecydowano o podziale dokumentu Bibliografii na oddzielne pliki, zawierające kolejne roczniki *Bibliografii...*. Tak przygotowane źródła przetworzono za pomocą formatu ConTEX – trzeba przyznać, że jest to dość duże obciążenie, zarówno dla kompilatora, jak i dla samego procesora. Podsumowaniem niech będzie kilka porównań (tab. 1).

Pełna 3-przebiegowa kompilacja całości dokumentu (trzy roczniki) trwa ok. 2 godziny 17 minut (szacunkowych pomiarów dokonano na komputerze Intel Pentium 4, 1.61 MHz, 512 MB RAM, system operacyjny Windows XP Professional).

PODZIĘKOWANIA: Autorzy artykułu składają podziękowania: Hansowi Hagenowi – za ConTEXt-a, Jurkowi Ludwichowskiemu i Pawłowi Jackowskiemu – za pomoc translatorską, oraz Markowi Wójtowiczowi i Tomkowi Dunin-Wąsowiczowi – za pomoc w realizacji aplikacji.

Literatura

[1] Swianiewicz J., *Pakiet MAK jako narzędzie do produkowania danych dla TEX*, Biuletyn GUST 16, s. 79–89.

[2] Bolek P., *ConTEXt – elegancja i efektywność*, Biuletyn GUST 15, s. 11–20.

[3] Hagen H., *TEX as presentation tool – an introduction to the ConTEXt presentation*, Biuletyn GUST 15, s. 57–61.

[4] Dokumentacja formatu ConTEXt, `http://www.pragma-ade.nl/`.

◇ Jacek Kmiecik
jk@agh.edu.pl

◇ Marek A. Valenta
valenta@agh.edu.pl

[3] W kolejnych latach, sukcesywnie przybywa jeden rocznik.