

Science and engineering databases in an Open Source software world

Radosław Urbaś

urbas@student.uci.agh.edu.pl

<http://student.uci.agh.edu.pl/~urbas/mownit>

Streszczenie

Praca opisuje systemy baz danych w projektach naukowych opierając się na kilku przykładach. Preferowane są narzędzia i systemy stanowiące Wolne Oprogramowanie. Rozdział pierwszy przedstawia pokrótce świat baz danych, ich historię i zastosowania. Rozdział drugi to krótkie wprowadzenie w ideę Open Source. Trzeci rozdział to ogólny opis rozwiązania zadanego przykładowego problemu, nie skupia się na szczegółach, a raczej na ogólnej, ewolucyjnej idei rozwiązania. W kolejnym rozdziale prezentowana jest baza danych służąca do zbierania i udostępniania informacji na temat wirusów, jest to istotne zagadnienie praktyczne - ze względu na masowe podróże po całym globie i szybkie rozprzestrzenianie się wirusów. W dzisiejszych czasach potrzeba dynamicznego rozwoju biotechnologii jest bardzo duża. Piąty rozdział - to prezentacja, na przykładzie Surface Science Spectra, centralnej bazy danych - która tworzona jest przez bardzo dużą rozproszoną grupę użytkowników. Dane z bazy tej także są szeroko udostępniane. Ostatni rozdział pokrótce przedstawia kilka narzędzi i technologii bazodanowych, stworzonych na podstawie idei Open Source.

1 Wstęp

Pierwotnym zastosowaniem aplikacji bazodanowych było przechowywanie i zarządzanie danymi pochodzącymi z badań naukowych oraz wszelakich informacji technicznych. Z czasem jednak bazy danych w coraz większym stopniu zaczęły być wykorzystywane w celach biznesowych. W szczególności duże komercyjne przedsiębiorstwa dużo mocniej niż inżynierowie i naukowcy wykorzystywały bazy danych. Rozwój systemów baz danych stał się zależny od potrzeb biznesu. Naukowcy i inżynierowie zostali pozostawieni sami sobie, mogli tworzyć i rozwijać własne systemy lub dostosowywać systemy biznesowe do swoich potrzeb.

Przełom w rozwoju systemów baz danych stanowił gwałtowny rozwój Internetu. Globalna sieć zmieniła kierunek dalszych prac nad bazami danych. E-biznes stał się bardzo szybko rozwijająca się dziedziną, co pociągało za sobą szybki rozwój baz danych - zorientowanych na współdziałanie z użytkownikiem poprzez sieć. Przeglądarka internetowa stała się kluczowym sposobem dostępu do danych. Dobrym przykładem są sklepy internetowe: to już nie tylko wyszukiwanie towarów i cen, ale także indywidualne podejście do klienta. Personalizacja zawartości strony internetowej pod kątem danego klienta. Informacje o odwiedzającym

przechowywane są w bazie, dzięki czemu przy kolejnej wizycie można prezentować mu potencjalnie najodpowiedniejsze (z biznesowego punktu widzenia) dla niego treści.

Taki właśnie wielodostęp i indywidualne podejście do użytkownika są naturalnymi mechanizmami potrzebnymi w czasie prowadzenia dużego projektu naukowego. Projekt skupia specjalistów wielu różnych dziedzin, różnorodne dane muszą być przechowywane i udostępniane wszystkim zainteresowanym, którzy najczęściej przebywają w różnych miejscach. Przechowywane dane to nie tylko tekst, ale wszelakie, także multimedialne obiekty. Łatwy i skuteczny dostęp skłonił projektantów i inżynierów do pracy nad wolnym oprogramowaniem udostępniającym takie funkcjonalności jak aplikacje komercyjne, w celu wykorzystania ich między innymi w projektach naukowych.

W ciągu ostatnich trzech dekad współdzielone naukowe bazy danych przeżyły wielką ewolucję, od tabel z liczbowymi wynikami po kolekcje rekordów zawierających najróżniejsze rodzaje danych. Papierowe notatki zostały zastąpione przez formę elektroniczną, która od razu może być wysłana np. poprzez email. Trzydzieści lat temu dane przekazywane były w postaci wydruków, bardzo długich tabel - możliwych do czytania przez ludzi, ale już nie przez maszyny. Jak już wcześniej wspomniano - kluczowym wydarzeniem, był rozwój Internetu. Globalna sieć otworzyła nowe drogi rozwoju. Opracowanie to ma na celu przedstawienie (na przykładach) rozwiązań stosowanych w projektach naukowych. Zarówno ogólnych idei i wzorców rozwiązań, jak i wskazania konkretnych narzędzi.

2 Open Source

2.1 Open Source Software

Oprogramowanie o otwartych źródłach, dostępne na podstawie wolnej licencji (najbardziej popularna GNU GPL ¹).

2.2 Open Science

Projekt stworzony w celu tworzenia naukowego oprogramowania Open Source. Znakomita część pracy naukowców zależy od narzędzi jakimi dysponują przy analizie danych eksperymentalnych i ich powiązań z modelami teoretycznymi. Dla większości projektów problemem nie jest już sam sprzęt komputerowy - gdyż w dzisiejszych czasach jest on osiągalny dla większości ludzi, ale brakującym elementem jest oprogramowanie umożliwiające pracę na nim. W rozwiązaniu tego problemu pomóc ma projekt Open Science ².

2.3 Open Access

Definiuje się jako wszechstronne źródło wiedzy i dziedzictwa kulturowego ludzi, informacje które zostały zaakceptowane społecznością naukową. Na udostępnia-

¹<http://www.gnu.org/licenses/gpl.html>

²<http://www.openscience.org/>

ne informacje składają się oryginalne wyniki badań, materiały źródłowe, cyfrowe obrazy, materiały multimedialne, artykuły, opracowania itp. Jedną z części jest np. Open Access Journals ³.

2.4 Open Biotechnology

Jest inicjatywą na rzecz rozszerzenia idei Open Source w biotechnologii i innych pokrewnych naukach. Celem jest umożliwienie badań w lokalnych ośrodkom, poprzez udostępnienie im rozwiązań, wyników uzyskanych przez innych, gdyż przynosi to obopólne korzyści. W tak ważnej w dzisiejszych czasach dziedzinie jaką są biotechnologie - wymiana informacji jest kluczem do sukcesu ^{4 5}.

3 Projekt Norman'a Chonacky i Dante Choi

3.1 Zadanie

Przykładem tworzenia systemu w oparciu o narzędzia open-source był projekt stworzony przy okazji Earth Systems Questions in Experimental Climate Change Science. W swoim artykule ([2]) Norman Chonacky i Dante Choi opisują wykorzystywane przez nich technologie.

Celem projektu było stworzenie witryny internetowej umożliwiającej prezentowanie notatek, nagrań z konferencji, obrazów wideo dyskusji plenerowych i tym podobnych. System miał być dostępny dla wszystkich uczestników seminarium. Rozwiązaniem problemu było stworzenie dynamicznej witryny internetowej. Ważnym aspektem projektu było to, że użytkownikami systemu mieli być specjaliści różnych dziedzin. Nie dysponując odpowiednim systemem, postanowili samodzielnie stworzyć system umożliwiający dostęp do tych różnorodnych typów danych. Dynamiczna strona internetowa stanowiła interfejs przez który można otrzymać rządane dane.

3.2 Rozwiązanie początkowe

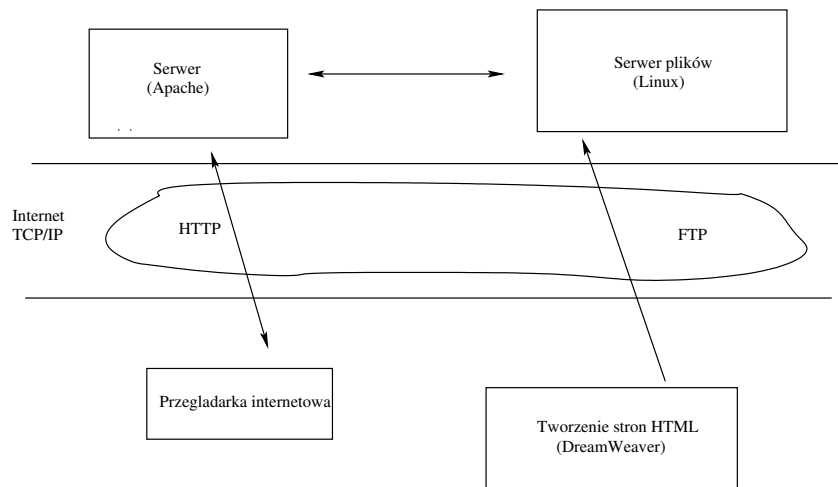
Z względu na specyfikację zadania zastosowano strukturę klient-serwer. Na serwer składają się zarówno hardware - czyli maszyna host'a, jak i jej oprogramowanie. W projekcie tym zastosowano Apache na Macintosh OS X. W najprostszej, statycznej konfiguracji aplikacja serwera nasłuchuje na zapytania wysyłane przez przeglądarkę klienta. Argumentem jest względna ścieżka do pliku jaki klient chce otrzymać, serwer odpowiada poprzez wysłanie rządanego pliku. Ten prosty system ilustruje poniższy rysunek.

Edytor HTML jest opcjonalny dla zwykłego edytora tekstu, aczkolwiek upraszcza i przyspiesza on prace nad stroną www. W swoim projekcie autorzy wykorzystali Macromedia Dreamweaver(komercyjna aplikacja, dostępna wersja shareware).

³<http://www.doaj.org/articles/about>

⁴<http://dmoz.org/Science/Biology/Biotechnology/>

⁵<http://www.cambia.org/>



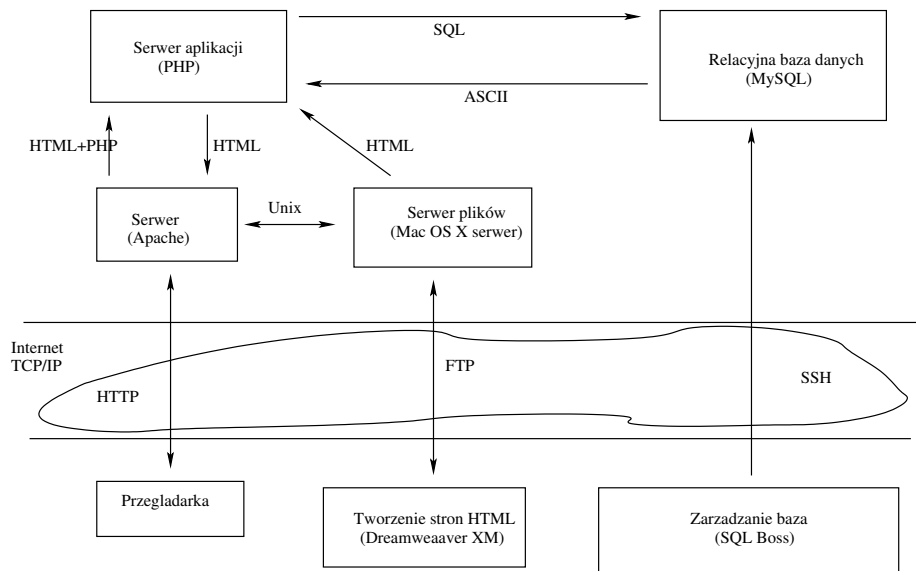
Rysunek 1: klient-serwer

3.3 Rozwinięte rozwiązanie

Aby umożliwić dynamiczne tworzenie stron HTML po otrzymaniu zapytania przez HTTP dodane zostały nowe komponenty. Poprzez skorzystanie z PHP dostajemy możliwość wykonanie zadania, które otrzymano od klienta. Umożliwia to przykładowo wstawienie do generowanej strony HTML danych z bazy, których zażyczył sobie klient i odesłanie ich właśnie w postaci strony www. Zastosowanie takiej technologii - PHP + MySQL - jest popularne ze względu na jej stosunkową prostotę i brak szczególnych wymagań po stronie klienta. Strona www - po stronie klienta jest generowana na bieżąco, zgodnie z zapytaniami jakie kieruje on do systemu. Interfejs klienta zazwyczaj jest graficzny, poprzez przyciski i pola tekstowe wprowadzane są zapytania.

Zasadniczą rzeczą która nas interesuje w tym projekcie jest sama baza danych. Autorzy próbowali najpierw baz dedykowanych do dostępu poprzez Internet, jednak ze względu na ograniczone możliwości i niezadawalające rozwiązania musieli z nich zrezygnować. Ostatecznie zdecydowano się na MySQL. Fakt, że źródła MySQL są otwarte - był dla autorów ważny, ze względu na wyznawaną filozofię otwartości.

Do manipulowania danymi w bazie MySQL użyto interfejsu dostarczanego przez SQL Batch Object Submission System. Narzędzie to było interesujące dla autorów ze względu na odpowiadający im graficzny interfejs, który generuje zapytania SQL wysyłane następnie przez Telnet lub SSH.



Rysunek 2: klient-serwer

MySQL:

1. bardzo popularny
2. wolny (na licencji GNU GPL, a także na licencji komercyjnej)
3. elastyczny
4. niski koszt wdrożenia

3.4 Podsumowanie

Zaproponowane rozwiązanie jest hybrydą Open Source i programów komercyjnych, stworzoną w celu rozwiązania zadanego problemu - stworzenia dynamicznej strony internetowej - przeznaczonej dla specjalistów różnych dziedzin, w celu przechowywania i wymiany danych różnych typów. Jakkolwiek nie jest optymalne, oraz wiąże się z pewnymi kosztami - spełnia wymagania aplikacji przeznaczonej do celów naukowych i inżynierskich. Zaletą jest inwestowanie i rozwijanie, na rzecz wspólnego dobra, w oprogramowania Open Source. Wadą - trudności w instalowaniu dla niedoświadczonych użytkowników.

4 The Universal Virus Database ICTVdB

4.1 Cel

Baza danych tworzona przez The International Committee on Taxonomy of Viruses stanowi uniwersalne narzędzie wspomagające zrozumienie zależności panujących pomiędzy wirusami. Podstawowym celem systemu jest precyzyjne identyfikowanie wirusów i dołączanie informacji na ich temat.

Projekt ICTVdB⁽⁶⁾ powstał w 1991 roku - jako jeden z pierwszych tego typu. Celem projektu jest stworzenie bazy zawierającej informacje dotyczące wszystkich wirusów zwierzęcych, roślinnych, bakterii, grzybów na wszystkich poziomach taksonomii. Początkowo rozwijany był na Australian National University⁽⁷⁾, obecnie na Columbia University's Biosphere 2 Center.

4.2 Taksonomia

Taksonomia bazuje przede wszystkim na morfologii organizmów. Pomimo tego, że obecnie wirusy można oglądać po mikroskopem elektronowym, to są one raczej rozpoznawane na podstawie ich cech chemicznych i genetycznych. Baza danych używa systemu Delta (Description Language for Taxonomy⁸), który jest światowym standardem wymiany danych w taksonomii. Przechowywane informacje rozciągają się od własności molekularnych po geograficzne pochodzenie. Tak różnorodne dane jak zależności pomiędzy genami, struktura białkowa, czy sposób infekcji są przechowywane w systemie Delta. Ponieważ nazwy wirusów zmienniają się często, zawierają nazwy geograficzne, ich identyfikowanie jest kłopotliwe. W tym celu wprowadzono kod analogiczny do nomenklatury panującej w badaniach nad enzymami. Kod przypomina adres IP pozwalający identyfikować komputery w sieci. Jego budowę ilustruje poniższa tabela, na przykładzie wirusa Polio.

Analizując kody łatwo stwierdzić relacje między dwoma wirusami. Przykładowo, pokrewieństwo dwóch członków rodziny *Reoviridae*, *Mal del Rio Cuarto virus* i *Nilaparvata lugens reovirus* jest natychmiast widoczne gdy spojrzymy

⁶<http://www.ncbi.nlm.nih.gov/ICTVdb/index.htm>

⁷<http://www.rsbs.anu.edu.au/index.asp>

⁸<http://delta-intkey.com/www/overview.htm>

Taxonomic level	Decimal code
Order	00.=(not assigned)
Family	00.052.=Picornaviridae
Subfamily	00.052.0.=(no subfamilies)
Genus	00.052.0.01.=Enterovirus
Subgenus(serogroup)	Superceded by species concept
Species(type speciec)	00.052.0.01.001.=Poliovirus
Species	00.052.0.01.007.=Poliovirus
Subspecies	00.052.0.01.007.00.=(not assigned)
Subtype	00.052.0.01.007.00.001.=Poliovirus 1
	00.052.0.01.007.00.002.=Poliovirus 2
	00.052.0.01.007.00.003.=Poliovirus 3
Isolate(strain)	00.052.0.01.007.00.001.001=PV-1 Mahony
	00.052.0.01.007.00.001.002=PV-1 Brunhile
	00.052.0.01.007.00.002.001=PV-2 Lansig
	00.052.0.01.007.00.003.001=PV-3 Leon/37

na ich kody: 00.060.0.07.004 i 00.060.0.07.008. Początkowe zera są nadmiarowe - wprowadzone, by zaspokoić potencjalny rozwój taksonomi i wprowadzenie wyższych kategorii. Obecnie na kod składa się 19 cyfr.

4.3 Baza danych

Wraz z wzrostem wielkości bazy zaletą takiego kodowania nie jest tylko to, że jest ono unikalne, ale także to, że kod stanowi nazwę pliku, stosowaną do transportu informacji z ICTVdB poprzez sieć, i wiele baz takich jak GenBank, European Molecular Biology Laboratory, czy Swiss-Port, stosuje ten właśnie kod jako numer dostępowy do ICTVdB. Taksonomiczne bazy takie jak Springer Index of Viruses także używają tego kodu w celu podłączenia do ICTVdB.

ICTVdB nie jest typową relacyjną bazą danych, stanowi raczej płaski system plików, wszystkie komponenty ICTVdB zapisane w formacie Delta stanowią czytelny dla człowieka plik tekstowy. Umożliwiona jest edycja plików poprzez specjalny edytor ułatwiający znacznie pracę. Baza przechowuje różnorodne dane takie jak: kształt wirusa, budowa chemiczna, czy też miejsce pochodzenia, a także komentarze czy rysunki. W rzeczywistości baza stanowi jedną dużą tabelę, która pozwala na elastyczność taką jak relacyjna baza danych, zachowując jednocześnie to, że relacje w taksonomii są liniowe. Ilustracje wirusów przechowywane w bazie są wykorzystywane na wiele sposobów. Po pierwsze opis morfologiczny jest bardziej precyzyjny jeśli dołączymy do niego obrazy pochodzące z mikroskopu. Co nie jest dziwne, obrazy wirusów są także najczęściej poszukiwaną w sieci informacją dotyczącą wirusów. Fizycznie pliki przechowywane są poza bazą, często nawet dostęp do nich jest realizowany poprzez Internet (przykładowo dane na temat *Norwalk virus* ⁹)

⁹<http://mmtsb.scripps.edu/viper/>

System Delta pozwala na łatwe pobieranie danych z bazy, generowanie raportów z danych w postaci np. strony HTML (przykład ¹⁰). Tworzone są także inne narzędzia dostępu, takie jak applety Java, skrypty pozwalające wyświetlić drzewa katalogów, pracuje się nad zastosowaniem XML. Standard Delta spełnia stawiane mu wymagania już od ponad trzydziestu lat.

5 Surface Science Spectra: A Hybrid Journal-Database

5.1 Opis zagadnienia

Spektroskopia to dziedzina zajmująca się badaniem oddziaływania promieniowanie na materię. Jest jedną z podstawowych metod stosowanych przez chemików, fizyków, w celu identyfikacji, rozpoznawania i badania materii. Szeroko stosowana w praktyce. Spektroskopia jest też często rozumiana jako ogólna nazwa wszelkich technik analitycznych polegających na generowaniu widm. Istnieją różne rodzaje spektroskopii, m.in.:

1. spektroskopia ramanowska
2. spektroskopia świetlna: UV, VIS i IR
3. spektroskopia rentgenowska
4. spektroskopia NMR
5. spektroskopia EPR
6. dichroizm kołowy
7. spektroskopia elektronowa
8. spektroskopia neutronowa
9. spektroskopia mas
10. spektroskopia sił atomowych
11. spektroskopia akustyczna

Wyniki zadań z wykorzystaniem spektroskopii - widma - przechowywane muszą być w specjalnych kolekcjach. Istnieje wiele przesłanek ku temu, aby zastosować bazy danych, w celu gromadzenia wyników. Zwykle konieczne jest przechowywanie licznych wzorców - w celu zidentyfikowania nieznanego materiału. Kolejnym ważnym powodem jest udostępnianie i wymiana informacji pomiędzy różnymi ośrodkami. Tradycyjne metody nie pozwalają na dokładne i szerodostępne prezentowanie wyników. Rozwój w tej dziedzinie rozpoczął się w 1990 roku gdy American Vacuum Society zajął się tworzeniem bazy danych widm, zwaną Surface Science Spectra⁽¹¹⁾. Przedsięwzięcie ma na celu stworzenie wysokiej jakości bazy, m.in. poprzez współpracę społeczności przy jej rozwoju. Użytkownicy, w szczególności studenci i nowi uczestnicy, tworząc nowe rekordy

¹⁰<http://delta-intkey.com/www/desc.htm>

¹¹<http://www2.avv.org/sss/>

są kontrolowani przez społeczność naukową. Ważną sprawą, jest masowy udział w tworzeniu wspólnego dobra, wykorzystywanego na całym świecie. Każdy może mieć satysfakcję, że przyczynił się do rozwoju tej gałęzi nauki.

W czasie rozwoju SSS, naprzód poszła technologia - Internet bardzo ułatwił komunikację i współpracę. Rada zarządzająca i inni biorący udział w tworzeniu systemu SSS musieli rozwijać i tworzyć nowe rozwiązania, by wykorzystać nowe możliwości. Te podjęte wyzwania skutkują tym, że zgromadzone dane są tym bardziej cenne dla każdego, kto zajmuje się spektroskopią. W roku 2002 ponad 350 osób (z 101 instytucji, z 19 krajów) współpracowało w tworzeniu bazy, a SSS-Online dostawało około 2000 zapytań o dane miesięcznie z ponad 400 miejsc.

5.2 Przechowywanie dane

Baza danych widm przechowuje więcej niż tylko po prostu pary współrzędnych (x, y) które określają widmo. Wiele innych czynników musi być opisanych, m.in.:

1. informacja o analizowanym materiale
2. rodzaj użytego spektrometru
3. warunki w jakich zostały uzyskane dane
4. procedura kalibracji
5. wyniki analizy eksperymentu
6. metody zastosowane w celu uzyskania wyników

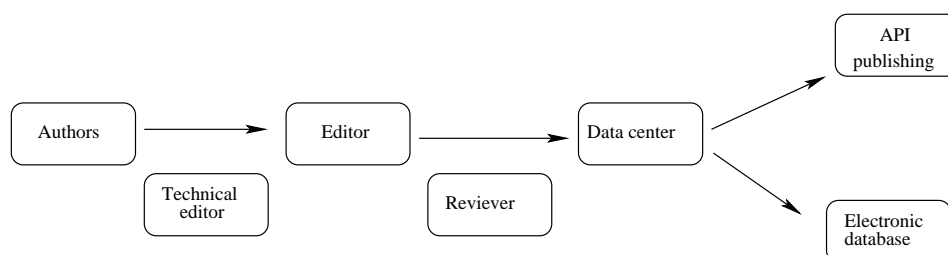
Aby dane były użyteczne dla innych, wiele dodatkowych parametrów musi zostać opisane (dana XPS/AES składa się z 188 elementów). Po wielu analizach ustalono standard uzyskiwania i redagowania danych. Kompletność danych jest istotna, gdyż będą wykorzystywane do celów, których nie może przewidzieć eksperymentator. Z drugiej jednak strony, jeśli zasadę tą będzie się stosować zbyt restrykcyjnie, wielu potencjalnych uczestników przedsięwzięcia nie będzie w stanie spełnić tych wymagań. Dlatego trzeba utrzymywać równowagę pomiędzy tymi dwoma rozbieżnymi celami.

Jakość danych jest ściśle kontrolowana, ze względu na ilość spełnionych warunków klasyfikuje się je w pięciu klasach, gdzie Level 1 - to rekordy, które bezwzględnie spełniają wszystkie wymagania, Level 2 - wszystkie warunki - z wyjątkiem sytuacji wyjątkowych, Level 3 i 4 - to zalecane i najczęściej występujące rekordy, Level 5 - wiele poszczególnych pól jest opcjonalnych. Posiadając taką klasyfikację obie strony mogą dokonać odpowiedniej selekcji danych.

Dane w bazie podzielone są na trzy klasy: reference - najwyższa jakość i stopień kompletności - dane uzyskane przy użyciu najlepszych dostępnych narzędzi, na najczystszych powierzchniach, comparison - najpopularniejsza, pomiary dokonane przy użyciu określonej klasy narzędzi, chemicznie określone powierzchnie, technical - słabo określone powierzchnie, ale potencjalnie ciekawe wyniki (skorodowane powierzchnie, powierzchnie ścian komórkowych itp.).

5.3 Obieg danych, oprogramowanie i format dokumentów

Autor danych kompletuje je i wysyła drogą elektroniczną w zadanym formacie do biura SSS. Następnie dane dostępne są w wersji papierowej, jak i poprzez witrynę internetową SSS. Dane otrzymane od współpracowników są sprawdzane. Po zaakceptowaniu ich na wszystkich szczeblach są wprowadzane do bazy. Baza danych SSS to relacyjna baza danych, Paradox. Po sformatowaniu danych (program Xyvison) artykuły przesyłane są do publikacji do American Institute of Physics (AIP). Format przesyłanych danych - to Postscript. AIP Online Journal Publishing Service (OJPS) udostępnia dane poprzez SSS-Online. Dane są udostępniane w formacie PDF (przykładowy dokument ¹²). Wszystkie zbiory mogą być przeszukiwane przy użyciu udostępnionych narzędzi¹³. Obieg danych ilustruje schemat:



Rysunek 3: obieg danych

5.4 Podsumowanie

Rewolucja w postaci elektronicznych dokumentów i rozwój sieci Internet miały zdecydowany wpływ na rozwój SSS. Na początku - nowy rekord miał postać kilku papierowych stron wpinanych do odpowiedniego segregatora. Obecnie zasoby SSS przechowywane są w komputerowej bazie danych, a dostępne są zdalnie, w przyjaznej dla użytkownika formie - poprzez automatyczną konwersję i generowanie przejrzystych dokumentów. Pracuje się nad interaktywnymi aplikacjami, które umożliwią dynamiczne przeglądanie interesujących użytkownika danych - skupianie się na wybranym fragmencie widma, jego powiększenie itp. Rozwój technologii bazodanowych wybitnie wspomógł ten projekt, umożliwiając działanie na dużo większą skalę - co daje natychmiast rezultat w postaci zwiększenia ilości posiadanych informacji i ułatwienia oraz przyspieszenia rozwoju wielu dziedzin nauki, które wykorzystują spektroskopię.

¹²<http://sss.avs.org/20020301.pdf>

¹³<http://scitation.aip.org/>

6 Przykładowe narzędzia Open Source

6.1 Serwery baz danych

6.1.1 Firebird

Firebird ¹⁴ to relacyjna baza danych (ANSI SQL-99) przeznaczona pod Linux, Windows, a także różne platformy Unix'owe. Firebird oferuje wysoką wydajność, i silne wsparcie dla procedur składowanych i triggerów. Jest używana pod wieloma nazwami od 1981 (rozwijana na podstawie uwolnionego kodu InterBase SQL 6.0 ¹⁵).

Firebird jest niezależnym projektem multi-platformowego systemu zarządzającego relacyjną bazą danych bazowanym na źródle kodu uwolnionego przez Inprise Corp (teraz znany jako Borland Software Corp) 25 Lipca 2000 pod InterBase Public License v.1.0.

Firebird jest całkowicie darmowy, wolny od rejestracji, licencji czy podatków. Może być swobodnie rozwijany do użytku z innymi aplikacjami, bez względu na to czy są wolne, czy też komercyjne.

6.1.2 MySQL

Serwer baz danych MySQL ¹⁶ jest najbardziej popularnym serwerem typu Open Source. Ponad sześć milionów użytkowników zainstalowało MySQL do wsparcia rozbudowanych witryn internetowych, czy innych ważnych projektów. Przykładami dużych firm używających tego serwera to są The Associated Press, Yahoo, NASA, Sabre Holdings, czy Suzuki.

MySQL jest atrakcyjną alternatywą dla kosztownych komercyjnych technologii. Może być rozwijany do własnych potrzeb, jego kod jest otwarty.

6.1.3 PostgreSQL

PostgreSQL ¹⁷ to obiektowo-relacyjny system zarządzania bazą danych (ORDBMS) bazowany na POSTGRES, Version 4.2, stworzonym na University of California, Berkeley Computer Science Department. POSTGRES był pionierski w wielu dziedzinach, które później zostały włączone do niektórych systemów komercyjnych

PostgreSQL jest Open Source'owym potomkiem tamtego oryginalnego systemu z Berkeley. Spełnia standardy SQL92 i SQL99 oraz oferuje wiele możliwości, takich jak: kwerendy, klucze obce, triggerzy, perspektywy, integralność transakcji. Dodatkowo PostgreSQL może być rozszerzany poprzez dodawanie nowych: typów danych, funkcji, operatorów, procedur składowanych itp.

¹⁴<http://www.firebirdsql.org/>

¹⁵<http://www.borland.com/interbase/>

¹⁶<http://www.mysql.com/>

¹⁷<http://www.postgresql.org/>

Ponieważ wydawany jest na wolnej licencji, PostgreSQL może być wykorzystywany, modyfikowany i dystrybuowany przez każdego bez ograniczeń. Zarówno przez osoby prywatne, komercyjne firmy, czy społeczność akademicką.

6.1.4 SQLite

SQLite ¹⁸ to niewielka biblioteka napisana w C która implementuje samodzielny, rozszerzalny, niewymagający konfiguracji silnik bazodanowy. Właściwości:

1. transakcje są atomowe i niezawodne nawet w przypadku awarii systemu
2. nie wymaga konfiguracji - nie potrzebna specjalna administracja
3. implementuje większość standardu SQL92
4. cała baza trzymana jest jako jeden plik dyskowy
5. nie ma problemów przy współpracy dwóch maszyn o różnej kolejności bitów (big i little endian)
6. obsługuje bazy o wielkości do 2 TB
7. mała wielkość - około 30000 lini kodu w C, 250 KB skompilowanego kodu
8. szybka i prosta w obsłudze
9. dobrze skomentowany kod
10. samodzielna - nie wymaga zewnętrznych odwołań
11. źródła są otwarte - dla wszelkiej działalności
12. dostarczana z specjalny programem dostępowy do bazy

6.2 Tworzenie baz danych

6.2.1 Kexi

Kexi ¹⁹ to zintegrowane środowisko służące do zarządzanie danymi. Wspomaga tworzenie schematu bazy, umieszczanie danych w bazie, zapytania i przetwarzanie danych. Potrzeba rozwoju tego typu aplikacji wynikła z braku Open Source'owego programu służącego do zadań realizowanych przez komercyjne aplikacje typu MS Access, FoxPro czy Oracle Forms.

6.2.2 Rekal

Rekal ²⁰ stanowi aplikację służącą do obsługi bazy danych, w stylu MS Access, z tym że Rekal sam w sobie nie zawiera bazy danych. Dane przechowywane są na serwerze SQL, a Rekal jest narzędziem do wydobywania, prezentowania i uaktualniania danych.

Większość wydana jest na licencji GPL. W zasadzie wszystko, poza sterownikami do komercyjnych baz danych. Sterowniki do MySQL, PostgreSQL, XBase/XBSQL i DBTCP są dołączone na licencji GPL.

¹⁸<http://www.sqlite.org/>

¹⁹<http://www.kexi-project.org/>

²⁰<http://www.rekallrevealed.org/>

6.2.3 JEDI Database Desktop

JEDI Database Desktop ²¹ jest niekomercyjnym Open Source'owym programem służącym do dostępu do bazy danych. Zastępuje on i rozszerza Borlands Database Desktop, który był komercyjną aplikacją. Wydany na Mozilla Public License, MPL i dostępny w wersji na 32-bitowe środowiska Windows (Linux'owy Kylix planowany w przyszłości).

Bibliografia

1. Andrés Guadamuz: Open Science: Open Source Software Licenses and Scientific Research *20th BILETA Conference: Over-Commoditised; Over-Centralised; Over-Observed: the New Digital Legal World? April, 2005, Queen's University of Belfast*
2. Norman Chonacky, Dante Choi: Science and Engineering Databases in an Open-Source Software World *Computing in science and engineering. May/June 2003 (Vol. 5, No. 3) pp. 10-13*
3. Cornelia Büchen-Osmond: The universal virus database ICTVdB *Computing in science and engineering. May/June 2003 (Vol. 5, No. 3) pp. 16-25*
4. Overview of the DELTA System: <http://delta-intkey.com/www/overview.htm>
5. National Center for Biotechnology Information: <http://www.ncbi.nlm.nih.gov/>
6. Wikipedia: <http://pl.wikipedia.org/wiki/Spektroskopia>
7. Stephen W. Gaarenstroom: Surface Science Spectra: A Hybrid Journal-Database *Computing in science and engineering. May/June 2003 (Vol. 5, No. 3) pp. 26-30*
8. Surface Science Spectra: <http://www2.avs.org/sss/>
9. Jeffrey D. Ullman, Jennifer Widom: A First Course in Database Systems *Prentice-Hall, December 2001 ISBN: 0130353000*

²¹<http://sourceforge.net/projects/jedidbd/>