

Bazy danych w życiu naukowym

Paweł Misiak
13 maja 1996
Computerworld

Uprawianie nauki to między innymi zajmowanie się informacją. Korzystanie z niej, przetwarzanie i wymiana stanowią istotną część pracy każdego naukowca. Bez dostępu do rozmaitych danych nie można uprawiać nauki.

W potocznym wyobrażeniu, utrwalonym obrazami literackimi i filmowymi, naukowiec to ktoś spędzający całe dni w bibliotekach, wertujący tysiące zadrukowanych stron i pilnie wynotowujący ważne informacje. W gabinecie naukowca piętrzą się stosy segregatorów pełnych notatek i kopii artykułów oraz wypełnione fiszkami pudełka. W istocie, praca z literaturą przedmiotu to jedna z najważniejszych umiejętności naukowca.

Przegląd literatury

Nauka to system gromadzenia i weryfikacji informacji, które powiązane konstrukcjami logicznymi stają się teoriami.

Toteż przed pracownikiem nauki staje zadanie gromadzenia nowych informacji, odkrywania nowych faktów, tworzenia nowych teorii, wyjaśniających więcej i lepiej niż dotychczas. Znamię nowości jest nieodłączną cechą wartościowej pracy naukowej. Nic więc dziwnego, że naukowiec musi poświęcić wiele czasu na zapoznanie się z dorobkiem innych badaczy w interesującej go dziedzinie. Badania prowadzone w większości światowych laboratoriów przyczyniają się do stałej rozbudowy gmachu ludzkiej wiedzy przez uzupełnianie drobnych, brakujących fragmentów, w oparciu na fundamentach i większych elementach już istniejących. W drukowanych artykułach naukowych ten stan rzeczy objawia się obszernymi spisami cytowanej literatury (zdarzają się teksty, w których lista referencji bibliograficznych przewyższa objętością właściwą treść pracy).

Częstym problemem w pracy naukowca jest uzmysłowienie sobie, czy ktoś już kiedyś tego samego nie robił? Pytanie to spędza czasem sen z powiek, zwłaszcza wtedy gdy dotyczy wielomiesięcznego wysiłku badawczego i poniesionych w związku z tym kosztów. Jeśli bowiem rzecz została już przez kogoś "zrobiona", problem rozwiązany, teoria opisana, to włożona praca jest w dużej mierze daremna, a pieniądze zmarnowane. Powtórzone wyniki nie są wiele

warte na rynku naukowym. Stąd konieczność ciągłego śledzenia bieżącej i przeglądania dawnej literatury przedmiotu.

Komputer pomaga

Na polu naukowej "walki o literaturę" nieocenioną pomoc niosą komputery. Wyposażone w bazy danych i odpowiednie oprogramowanie do ich przeszukiwania, pozwalają na prowadzenie badań "literackich" w zakresie trudnym do wyobrażenia przy zastosowaniu metod tradycyjnych - wertowania czasopism, książek itp.

Kilkanaście lat temu, kiedy zaczynałem pracę naukową, spędzałem wiele godzin tygodniowo, przeglądając w bibliotece zeszyciki Current Contents. Wydawnictwo to zawiera spisy treści czasopism i innych wydawnictw periodycznych oraz książek z dziedziny nauk ścisłych, przyrodniczych i medycznych. Stanowi nieocenione źródło informacji na temat bieżącej produkcji naukowej. Niewiele jest na świecie bibliotek mających wszystkie ujęte w Current Contents tytuły. Wydawnictwo to pozwala zlokalizować ciekawe prace (głównie na podstawie tytułów i nazwisk autorów), a dzięki podaniu adresów autorów ułatwia zamówienie kopii, jeśli dany tekst nie jest osiągalny w którejś z pobliskich bibliotek. W sytuacji, gdy praca ukazała się w jakimś egzotycznym piśmie, jest to często jedyną możliwością dotarcia do niej.

Gdy kilka lat temu pojawiła się komputerowa wersja Current Contents, w postaci bazy danych na dyskietkach, życie stało się łatwiejsze. Zamiast żmudnego przeglądania większości tytułów publikacji w poszukiwaniu określonych treści, wystarczy podać programowi odpowiednie słowa kluczowe, by po kilku chwilach zobaczyć na ekranie spis prac mających związek z szukany profilem. Dyskietkowa wersja Current Contents ma tę przewagę nad drukowaną, że zawiera także streszczenia większości prac, dzięki czemu można precyzyjniej, niż tylko na podstawie tytułu, ocenić ich przydatność.

Program obsługujący Current Contents pozwala nie tylko przeszukiwać bazę według zadanego profilu, ale robić automatyczne "notatki", czyli zapisywać wybrane przez użytkownika rekordy w żądanym formacie na dysku albo od razu je drukować.

Pozwala też drukować gotowe prośby o przysłanie kopii tekstu wraz z nalepkami adresowymi autorów. Elektroniczna wersja Current

Contents przyspiesza pracę z literaturą przedmiotu, oszczędza wiele czasu i wysiłku.

Oprócz Current Contents, publikującego abstrakty prac naukowych jeszcze przed ich ukazaniem się w druku, istnieje wiele innych bibliograficznych, zazwyczaj retrospektywnych baz danych. Obejmują one różne gałęzie nauk: techniczne, medyczne, ścisłe, społeczne itd. Najszerzy bodaj zakres ma Science Citation Index. Rozprowadzany na dyskach optycznych, uaktualniany kwartalnie, pozwala prowadzić wielokierunkowe poszukiwania bibliograficzne. Podobnie jak we wspomnianym Current Contents, można tu podać programowi profil przeszukiwania w postaci słów kluczowych. Oprócz wyszukanych według podanego profilu prac można także "wyciągnąć" z bazy dane dotyczące prac cytowanych w każdym z wyszukanych artykułów, co przy umiejętnym posługiwaniu się bazą pomaga szybko zorientować się w całej literaturze omawiającej dane zagadnienie.

Bibliografie i cytowania

Bibliograficzne bazy danych w postaci elektronicznej pozwalają szybko i niemal bez wysiłku tworzyć prywatne zestawienia odnośników dotyczących literatury na dowolne tematy. Rodzi to pokusę zamieszczania w pisanych pracach sążnistych spisów referencji, z których nie wszystkie pozycje zostały przez cytującego - łagodnie rzecz ujmując - w całości przeczytane. Zdarzają się więc przypadki (można by rzec "wpadki"), gdy odnajdując pracę, na którą powołał się ten czy ów autor, stwierdzamy, że nie ma ona wiele wspólnego z tematem, poza przypadkową zbieżnością słów w tytule lub streszczeniu.

Innym ciekawym, praktycznym acz pozamerytorycznym zastosowaniem Science Citation Index jest wyszukiwanie w nim informacji, służących do ustalenia rankingu uczonych, np. do oceny jakości ich pracy. Wystarczy wyłować z bazy liczbę publikacji badanego oraz liczbę cytowań każdej z jego publikacji, pomnożyć je przez odpowiednie współczynniki i dodać do siebie. W ten sposób można dowieść liczbowo, że profesor Y jest 1,357 razy lepszy pod względem naukowym od docenta X. Oczywiście, diabeł tkwi w szczegółach, np. w doborze współczynników. Zawsze można ustawić je tak, by nasz instytut czy zakład okazał się najlepszy.

Fakty, wyniki, liczby

Literatura naukowa relacjonuje wyniki i metody prowadzonych badań. Opisuje fakty, teorie, podaje wyniki obserwacji i eksperymentów. Konstrukcje teoretyczne, podobnie jak zaobserwowane zjawiska, ważne z naukowego punktu widzenia, opisywane są w artykułach i książkach. Istnieje jednak ogromna liczba danych, zgromadzonych w wyniku obserwacji, eksperymentów i różnorodnych pomiarów, których publikowanie w postaci drukowanej nie ma sensu, jako że potrzebne są rzadko i zazwyczaj tylko wybiórczo.

W sytuacji, gdy w wielu laboratoriach na świecie rejestrowane są miliony faktów naukowych, najbardziej efektywnym sposobem gromadzenia i korzystania z nich jest tworzenie wielkich komputerowych baz danych. Nazwijmy je merytorycznymi. Dzięki sieci komputerowej i odpowiedniemu oprogramowaniu możliwa jest automatyzacja zarówno procesu zapisu danych, jak i szerokiego do nich dostępu (patrz: "Po co naukowcom Internet", CW nr 12/96).

W ten sposób powstają, są uzupełniane i udostępniane np. dane dotyczące genetyki muszki owocówki (*Drosophila melanogaster*), myszy, a nawet człowieka. Gromadzone informacje są dodawane do globalnej bazy danych i kopiowane na wielu serwerach internetowych. Dzięki temu badacz, zajmujący się problemami genetyki, nie musi przeszukiwać stosów zadrukowanego papieru, lecz formułuje proste zapytanie do bazy danych i po krótkim czasie otrzymuje wielostronną informację na dany temat.

Podobnie jest z danymi dotyczącymi ważnych biologicznie molekuł, jak białka czy kwasy nukleinowe. Istnieje kilka baz danych o zasięgu światowym, jak np. Protein Data Bank, gromadzących wyniki badań o strukturach i własnościach takich cząsteczek.

Przy okazji tworzenia rozmaitych baz opracowywane są też często standardy zapisu danych, co z kolei wpływa na szerokie i ujednolicone ich wykorzystanie do różnych celów. Znajduje odbicie w tworzonej oprogramowaniu, korzystającym z owych danych do analizy, wizualizacji czy dalszego przetwarzania. Na przykład większość programów do analizy i przetwarzania danych molekularnych wykorzystuje format stosowany w Protein Data Bank.

Duże bazy danych i odpowiednie oprogramowanie wspomagające ich wykorzystanie tworzone są także w innych dziedzinach nauki: fizyce

ciała stałego (np. krystalograficzne), w fizyce wysokich energii i cząstek elementarnych (wyniki eksperymentów akceleratorowych), geofizyce, tektonice, astronomii, a także w naukach inżynierskich (bazy danych materiałowe) i medycznych.

Za osobną klasę merytorycznych baz danych można uznać zbiory elektronicznych zapisów tekstów literackich i filozoficznych, gromadzone w ramach kilku szeroko zakrojonych inicjatyw, w rodzaju Projektu Gutenberg. Być może określenie "baza danych" jest tu nieco na wyrost, lecz znaczenie, jakie ma bezpośredni, szybki dostęp on-line do kilkuset tekstów światowego dziedzictwa kultury piśmienniczej (szkoda, że głównie anglojęzycznej), otwiera nowe perspektywy przed badaczami reprezentującymi niektóre gałęzie nauk humanistycznych. Łatwo je można sobie wyobrazić jeśli chodzi o badanie tekstów metodami informatycznymi, podobnymi do tych, jakie stosowane są przy analizie wielkich baz danych pod względem ilościowym. Zbiory te mają także, rzecz jasna, ogromne znaczenie dla zwykłego "zjadacza Internetu", dając mu bezpośredni dostęp do wielu klasycznych tekstów, których być może z biblioteki nigdy by sam nie wypożyczył.

Nauka - bazy - biznes

Różne bazy danych są w różny sposób udostępniane. W dziedzinach nauk podstawowych dostęp jest zwykle bezpłatny, często za pośrednictwem Internetu, lub prawie bezpłatny, np. za cenę nośnika CD-ROM. Tak jest w przypadku wspomnianych baz genetycznych i biomolekularnych. Jednak gdy dana baza ma także znaczenie komercyjne, czasem tylko w dalszym planie, jak w przypadku baz medycznych, bibliograficznych (sic!) czy ekonomicznych, chętni do skorzystania ze zgromadzonych w nich informacji muszą szeroko otworzyć portfele. Tworzenie, aktualizacja i oprogramowanie naukowych baz danych to nie od dzisiaj znakomity interes, dostęp do nich staje się bowiem w wielu dziedzinach codzienną koniecznością.