

Numerikus adatbázisok a tudományban és a technikában

A numerikus adatbázisok helye

A tudományos és műszaki numerikus adatbázisok nem olyan népszerűek az online keresők körében, mint a gazdasági numerikus adatbázisok. Az utóbbi években azonban ez a szakterület is jelentős fejlődésen ment át. 1975-ben csak 51 numerikus/faktografikus adatbázis volt Európában, 1983-ban a számuk már elérte az 1083-at.

Az adatbázisok különféle típusai: bibliográfiai, szöveges, numerikus stb. Kétféle csoportosításuk ismeretes; *Williams* [1] és *Cuadra* [2] az online adatbázisok alábbi kategóriáit különbözteti meg:

Williams	Cuadra
A) Szövegorientált adatbázisok: bibliográfiai teljes szövegű útmutató típusok	Hivatkozó adatbázisok: bibliográfiai forrás-tájékoztató típusok
B) Numerikus adatbázisok: tulajdonságok idősorok statisztikai információk	Forrásadatbázisok: numerikus szöveges-numerikus teljes szövegű szoftveradatbázisok

A numerikus és faktografikus (tény-)adatbázisokat gyakran összekeverik. Numerikus adatbázisok azok, amelyekben döntő többségben számadatok vannak, a hozzájuk szükséges minimális szöveggel. (Ezeket szokás adatbankoknak is nevezni, de a különféle szóhasználat miatti félreértések elkerülése érdekében célszerűbb a magyarban az egységes *adatbázis* szó használata, esetleg valamilyen jelzővel. – A ref.) A tényadatbázisokra elsősorban a szöveges tartalom jellemző.

A tudományos és műszaki numerikus adatbázisok elérhetősége és a használatuk iránti igények gyorsan nőnek, pedig az ilyen adatbázisok nem is annyira újak. A számítástechnika tért hódít a kutatásban és fejlesztésben, a laboratóriumi munkában és a tervezőasztalon; ezzel párhuzamosan fejlődik a numerikus/tényadatok számítógépes kereshetősége iránti igény.

A tudományos és műszaki életben minden bővebb és szűkebb szakterületnek más-más sajátosságai vannak. A *Committee on Data for Science and Technology (CODATA)* az adatokat a megszerzés módja szerint jellemzi:

- ▶ ismételhető mérések jól definiálható rendszerekben,
- ▶ megfigyelési adatok,
- ▶ statisztikai adatok.

Nemzetközi perspektívák

A tudományos és műszaki kutatás és fejlesztés nem áll meg az országhatárokon. Az USA technikája az elmúlt néhány évben világszerte tért hódított

külföldi kutató-fejlesztő központok és termelőüzemek révén. A szövetségi kormány által támogatott programok, mint pl. a fúziós energia kutatása, a nagyenergiájú fizikai jelenségek kutatása, nemzetközi kooperációban is folynak. Nemzetközi programnak számít a föld éghajlatának modellezése, az ökológiai kérdések megoldása stb.

A tudomány egyes területeinek országhatárokon túlmenő társadalmi vonatkozásai vannak (pl. az AIDS-vírus kutatása). 1987-ben USA–NSZK kooperációban indították el az AIDS adatbázist, hogy 30 olyan AIDS-vírus mutációt kövessék nyomon, amelyeket 1976 és 1986 között izoláltak.

A numerikus adatbázisok létrehozására irányuló együttműködési programok száma is gyorsan nő. Ide tartozó programok:

- ▶ a geofizikai adatok világ-adatközpontja,
- ▶ országos nukleáris adatközpont (NNDC),
- ▶ az éghajlatadatok számítóközpontja (CLICOM) stb.

Más programok az USA-n kívüli adatokból is táplálkoznak, mint az International Coal Database (nemzetközi szénbányászati adatbázis).

Szabványosítás

A numerikus adatbázisok nemzetközi terjedésével együtt merül fel a szabványosítás, az adatok egységesítésének kérdése. E területen a munka országos és nemzetközi szinten folyik. A CODATA 1966 óta foglalkozik az érvényes adatok ábrázolásának módszereivel [3].

Az anyagtulajdonságok adatbázisainak fejlesztése jól bizonyítja a szabványok szükségességét. Egyes esetekben ugyanis nagyszámú adatmező szükséges egy anyag azonosítására. Ehhez járul a nem szabványos terminológia és a mértékegységek gondja. Számos terület szabványosítására van szükség (az anyagok megnevezése, a tulajdonságok terminológiája és definíciója, mérési módszerek stb.).

A numerikus adatok szabványosításának kérdésével több országos és nemzetközi program, bizottság és intézmény foglalkozik. 1987-ben ennek a kérdésnek egy külön szimpóziumot szentelt az ASTM (American Society for Testing and Measurement = amerikai méréstechnikai társulat), amelyet 1989-ben egy másik követett. A CODATA is hasonló szimpóziumot tervez 1990-ben.

Az anyagtulajdonságok szabványosítási problémáján kívül hasonló munka folyik pl. a fizikában és kémiában is.

Az online rendszerek és adatbázisok

A két legfontosabb online keresési rendszer a *parancsnyelven alapuló* és a *menürendszerű*. Az utóbbit főleg a végfelhasználók kedvelik, amint azt nagy kutatóintézetek (pl. a Sandia Laboratórium)

munkatársai esetében tapasztaltuk. Van olyan adatbázis, amely menü- és parancsnyelvrendszerrel is kereshető, ilyenkor azonban a keresés hatékonyságán vitatkozni lehet.

Az általános használatra szánt vagy hálózatban hozzáférhető online rendszereknek felhasználóbarátnak kell lenniük. A tulajdonságokat a felhasználó számára ismert tárgyszavakkal kell kifejezni, vagy az értékeket a felhasználó által ismert tárgyszavakká átalakító programot kell alkalmazni. Erre a kérdésre egy felelet a metaadatok koncepciója, vagy az automatikus adattezaurusz elve [4].

A műszaki és tudományos életben legfontosabb *numerikus/faktografikus adatbázisok* jegyzéke az alábbiakban található, szolgáltatóközpontként csoportosítva.

BRS

EMIS (elektronikus anyagokra vonatkozó adatok, információk)

Hazardline (kémiai tulajdonságok, toxikológia, hulladékkezelés)

ICPR (integrált áramkörök paramétereinek keresése)

INSPEC (1987 óta numerikus és kémiai indexelés)

ISMS (ipari szabványok és katonai specifikációk)

Kirk-Othmer (kémiai tulajdonságok)

Merck Index (gyógyszerek)

SCPR (félvezetők paramétereinek keresése)

Dialog

Beilstein (kémiai tulajdonságok, szerkezet és nomenklatúra)

INSPEC (tulajdonságok 1987 óta)

Pascal (franciául, németül és angolul)

Kirk-Othmer Online (táblázatok kereshetők)

HEILBRON (fizikai és kémiai tulajdonságok, ábrák megjelenítése)

Merck Index (gyógyszerek)

Pergamon Orbit InfoLine

CHEMQUEST (termékek és forgalmazók – vegyipar)

Corrosion (korrózióállósági adatok)

Metals Data File (fémtechnikai és -fizikai adatok)

INSPEC

Pesticide Databank (kémiai tulajdonságok – növényvédő szerek)

STN International

Beilstein

CHEMLIST (toxikus vegyi anyagok jegyzéke)

CSCHEM (vegyianyag-információk)

CSEARCH-NMR (az STN-en keresztül hozzáférhető NMR spektrális adatok)

DIPPR (fizikaitulajdonság-adatok tervezőintézete)

INSPEC

JANAF (kritikailag értékelt kémiai-termodinamikai tulajdonságok)

NBS THERMO (termodinamikai tulajdonságok – szerves és szervetlen anyagok)

CIS

Aquatic Information Retrieval (vizek tudománya, toxikológia)

CESARS (kémiai értékelő- és keresőrendszer)

Chemical Carcinogenesis Research (orvosbiológia, toxikológia)

CHRIS (kémiai veszélyek elhárításának információs rendszere)

CNMR (szén-13 mágneses magrezonancia-adatok keresőrendszere)

CTCP (kereskedelmileg forgalmazott termékek klinikai toxikológiája)

Dermal Absorption (vegyi anyagok egészségre gyakorolt hatása)

Environmental Fate (vegyi anyagok szállítása és bomlása)

Genetic Toxicity (mutagenitás információk)

ISHOW (vízben levő veszélyes szerves anyagok információs rendszere)

IRSS (infravörös spektrumok keresőrendszere)

MSSS (tömegspektrumok keresőrendszere)

OHM-TADS (olaj és veszélyes anyagok – műszaki elhárítási adatrendszer)

PHYTOTOX (toxikológia)

RTECS (vegyi anyagok toxikus hatásainak nyilvántartása)

SANSS (kémiai szubstruktúra- és nomenklatúra-kereső rendszer)

Suspect Chemicals Sourcebook (vegyi anyagokra vonatkozó szabályozások és elővigyázatossági tudnivalók)

THERMO (szerves és szervetlen anyagok termodinamikai adatai)

WMSSS (Wiley-féle tömegspektrum-kereső rendszer).

Államilag támogatott rendszerek az USA-ban

Az USA szövetségi kormánya jóvoltából a fentiekén kívül évek óta működnek numerikus adatbázisok. Ezek közül több nyilvánosan is elérhető. Fontos példa az Energiaügyi Minisztérium (DOE) Országos Nukleáris Adatközpontjának (NNDC) öt nukleáris tárgyú numerikus adatbázisa. Az Országos Tudományos és Műszaki Intézet (NIST) 16 faktografikus/numerikus adatbázist állít elő és bocsát a kutatók rendelkezésére. Az Országos Orvosi Könyvtár (NLM) a toxikológiai információk kitűnő forrása.

Az állami hivatalok ezekhez az adatbázisokhoz közvetlenül vagy kommerciális szolgáltatókon keresztül biztosítanak hozzáférést. (A tapasztalat szerint az USA-n kívül az online hozzáférés az ilyen adatbázisokhoz többnyire erősen korlátozva van. – A ref.) Ennek ellenére mind nagyobb figyelmet fordítanak ezeknek az erőforrásoknak az eddigieknél jobb menedzselésére, hogy minél több kutatási-fejlesztési költség legyen megtakarítható. A DOE kutatói első lépésként összeállították a numerikus/faktografikus adatbázisok online útmutatóját. Másik példa az NTIS (Országos Műszaki Információs Rendszer) számító-

gépés adatállomány-mutatója, amely mintegy 50 USA-beli szövetségi hivatal, intézmény több mint 1300 értékes numerikus és szöveges adatállományát tartalmazza.

Mi várható a jövőben?

Az online ipar olyan újdonsága, mint a CD-ROM technika, várhatóan az adatbázisok online forgalmára hat vissza. CD-ROM-on fogják használni a ritkán aktualizálható adatbázisokat.

Az adatátviteli protokollok, az adatok és adatbázisok minőségének megbízhatósága és szabványosítása nélkül a hálózatok, a letöltés és a szakértő rendszerek fejlesztése nem következhet be. Fontos, mégis gyakran elhanyagolt tényező az ezeknek az adatbázisoknak a kereséséhez elengedhetetlen szakértelem. Az információközvetítők azzal segíthetnek felhasználóiknak és maguknak is, hogy a rendelkezésre álló numerikus/faktografikus adatbázisokat megismerik, és felmérik a várható igényeket. A

munkájuk jövőjét féltő online keresőknek ez esélyt jelent arra, hogy tudásukat erőforrásnak tekintsék, s ne mellőzzék őket.

Irodalom

- [1] WILLIAMS, M. E.: Computer-readable databases. 1. köt. Science, technology, medicine. Elsevier, 1985. p. VII–VIII.
- [2] Directory of Online Databases. Cuadra/Elsevier, 9. köt. 1. sz. 1988.
- [3] CODATA. Inventory of data sources in science and technology: a preliminary survey. UNESCO, 1982.
- [4] McCARTY, J.: The automated data thesaurus: a new tool for scientific information. Proceedings of the 11th International CODATA Conference. FDR, Karlsruhe, 1988.

/ALLAN, F. C.– FERREL, W. R.: Numeric databases in science and technology: An overview. = Database, 12. köt. 3. sz. 1989. p. 50–58./

(Roboz Péter)

Teljes szövegű online adatbázisok értékelése

Az utóbbi tíz évben az adatbázisok számának rohamos növekedésével arányosan nőtt a teljes szövegű adatbázisok száma is. Használatuk azonban nem nőtt olyan gyorsan, mint a bibliográfiai adatbázisoké. Az 1970-es évek végén három pontban állapították meg a tényleges használat feltételeit:

- ▶ az információ elegendő legyen,
- ▶ időben érkezzen,
- ▶ fontos legyen a felhasználónak.

Ezek közül legalább két feltételnek kell teljesülnie ahhoz, hogy egy adatbázis sikeres legyen. Az *American Chemical Society (ACS)* szerint teljes szövegű adatbázisaira mindhárom követelmény teljesül. Az ACS 1983-ban 18 folyóiratát teljes szövegű adatbázisok formájában tette hozzáférhetővé a BRS szolgáltatóközpontnál. Ezek az adatbázisok a *Chemical Journals Online (CJO)* adatbázis részeként kerültek át 1986-ban az STN International szolgáltatóközponthoz (USA–NSZK–Japán).

Az ACS az idők folyamán egyre több folyóirat-kiadó termékeit vitte be online kereshető formában a CJO adatbázisba. Ennek eredménye, hogy a teljes szövegű információ mennyisége valóban hatalmas lett: folyóiratcikkek ezreit tartalmazza az adatbázis, némelykor nyolc évre visszamenőleg. Az információ valóban friss, ugyanis a CJO-ban szereplő folyóiratok sokszor előbb olvashatók online, mint ahogy az előfizetőkhez a lap megérkezik. Végül a teljes szövegűben online kereshető és olvasható primer folyóiratok – a felhasználók szerint – a kémiai kutatás nélkülözhetetlen segédeszközei lettek.

A teljes szövegű adatbázisok előállítására

A teljes szövegű adatbázis előállításának két kulcskérdése van: a nyomtatott folyóiratok átalakítása online kereshető alakra és a nagy adatmennyiség tárolása.

A nyomtatott folyóiratok előállításában használt mágnesszalagos állományokat adatbázis-formátumra kell konvertálni, ami számos részproblémát rejt magában (pl. a nem alfanumerikus karakterek kódolása). Az egyes kiadók másképp kódolják az adatelemeket, ezért azokat az adatbázis-készítés során egységesíteniük kell, ami megnöveli a költségeket.

A betöltés után a tárolás költségei igen nagyok, mert teljes szövegeket, azaz minden egyes szót és a proximity (helyzeti operátoros) kereséshez szükséges indexeket is online kereshetővé kell tenni. A CJO adatbázisban az indexállományok tárolásához háromszor akkora tárterület szükséges, mint magához a szöveghez. Ha még grafikus információt is tárolni kellene, a költségek megengedhetetlenül nagyok lennének. A tárolás költségei így is sokkal nagyobbak, mint a bibliográfiai adatbázisok esetében.

A felhasználók

A felhasználók döntését, hogy teljes szövegű adatbázist használjanak-e, sok tényező befolyásolja. Az egyik ilyen a *teljesség* szempontja.