

Társadalomtudományi folyóiratcikkek visszakeresése

Az ISI (*Institute for Scientific Information = Tudományos Tájékoztatási Intézet*) bejelentette, hogy on-line hozzáférést biztosít a társadalomtudományi folyóiratcikkek legnagyobb információs tárához, a *Lockheed Információs Rendszerhez*. A rendszer *SOCIAL SCISEARCH* néven 1974 szeptemberétől működik. Az adattár 1200 társadalomtudományi folyóirat összes cikkére, továbbá 2500 természettudományi folyóirattól válogatott társadalomtudományi tárgyú cikkére terjed ki. Az adatbázis, amelyet 1972 óta gyűjtenek kb. 300 ezer tételt tartalmaz; havi gyarapodása mintegy 7000 tétel. A keresés többféleképpen lehetséges: kulcsszó, szótó, szóösszetétel, szerző és intézmény szerint. Hivatkozások szerint is lehet keresni. A *DIALOG* elnevezésű *Lockheed* program interaktív keresést tesz lehetővé. A Kaliforniában működő számítógéphez a terminálok telefonvonalon kapcsolhatók. Minden kikeresett tételről teljes bibliográfiai leírást kap az igénylő, kiegészítve a cikkekben foglalt hivatkozások teljes jegyzékével.

Journal of Library Automation, 8. k. 1. sz. 1975. p. 75–77./

(Sárdy Péter)

* * *

Konferencia a nagyméretű adatbázisok témakörében (USA)

1971-ben az *Egyesült Államok Tudományos Akadémiája* (*National Academy of Sciences*) kutatócsoportot hozott létre nagyméretű adatbázisok vizsgálatára. Bár a csoport eredetileg a *Kémiai Információs Bizottság* (*Committee on Chemical Information*) keretében alakult meg, a vizsgálatot később nem kémiai jellegű adatbázisokra is kiterjesztették.

A csoport meghatározása szerint *nagyméretű adatbázis*

nagyszámú rekordot, sokféle adatmezőt vagy adatelemet tartalmaz;

lehet nagyméretű azért mert nagyszámú kikereshető, kiválasztható elem van benne;

végül lehet nagy abban a könnyen érthető értelemben is, hogy nagy tároló kapacitást kíván.

A vizsgálat során csak olyan gépileg tárolt adatbázisokra szorítkoztak, amelyek legalább 100 millió karaktert tartalmaznak. Hamarosan kiderült, hogy a tárolás és

információkeresés alapelvei, fejlődési iránya és problémái nagyon hasonlóak a különböző alkalmazási területeken. Ez tette lehetővé és egyben szükségessé az alábbiakban ismertetett első konferencia megtartását, amelyre az USA Tudományos Akadémiáján került sor 1974. május 22–23. között a *National Research Council* (*Nemzeti Kutatási Tanács*) védnöksége alatt.

Nagy file-ok on-line működési módban

A konferencián tartott előadások közül sok foglalkozott az on-line működésű bibliográfiai információkereső rendszerekkel.

Az on-line szolgáltatás sikerét több tényező magyarázza. Egyrészt nőtt a retrospektív irodalomkutatás számára elérhető adatok mennyisége. Így például a *Chemical Abstracts* alapján, amely 1970-ben még csak két év anyagát tárolta géppel olvasható formában. 1974-ben már öt évre visszamenő irodalomkutatást lehetett elvégezni. Ugyanakkor növekedett az on-line szolgáltatás ismerőinek és felhasználóinak száma, fokozódott a felhasználók bizalma. A megnövekedett igény egyben gazdaságosabbá tette a nagyméretű adatbázisok fenntartását és üzemeltetését. Hasonló irányba hatott a technika fejlődése is, különösen a távközlési kommunikációs hálózatok kiépülése. Az on-line szolgáltatás ma már nem kuriózum, hanem kereskedelmi realitás. Becslések szerint 1974-ben 700 000 on-line irodalomkutatásra került sor, 1975-re számuk várhatóan egy millióra nő.

A nagyméretű adatbázisok problémái

A nagy és kis file-oknál egyaránt alapvető fontosságú a biztonságos adatkarbantartás. A véletlen és szándékos adatkárosodás elleni védekezéshez ismerni kell az adattárakat fenyegető veszélyeket.

Másik alapvető fontosságú probléma az adatok ábrázolásához felhasználható *karakterkészlet adekvátsága*. A kémiában és a biológiában kb. 1000 különböző karaktert használnak, más diszciplínák is hasonló számú karaktert igényelnek. Az egyes diszciplínák karakterkészletei azonban nem teljesen fedik át egymást. Ugyanakkor kimutatható az is, hogy például a kémiai szövegekben az összes előfordulás 94,5%-ában 36 alapjel szerepel (figyelmetlenül kívül hagyva a jel előfordulásának körülményeit: index, kisbetű, nagybetű stb.). A rendszerben használható karakterkészlet kialakításánál azonban figyelembe kell venni az input/output egységekkel és más file-okkal, szabványokkal való kompatibilitás követelményeit is.

A kémiai információkat és kereső rendszerekben nehezen megoldható feladatot jelent a kémiai vegyületek egyértelmű megnevezése, azonosítása is.

A tapasztalatok szerint az információs rendszer szűk keresztmetszete minden esetben az input adat rögzítés és

a „hard copy” (papírmásolat) output.

Az adatbázis nagysága problémát okoz, ha a rendszerben nincs olyan nagykapacitású tároló egység, amelyben az egész file elférne, ha bizonyos file-okat nem lehet invertálni, ha az indexelés inkonzisztens, ha a formátumok sőt a kódolás módja változott az évek során, ha nagyon sok változó írja le a komplex irodalomkutatás témáját, ha a file hierarchikus szervezésű stb.

Gazdaságosság

Becslések szerint az on-line működtetéshez szükséges inverz tárolás kb. százszor költségesebb, mint a mágnesszalagos. A nagyméretű adatbázisokkal kapcsolatos minden más manipuláció (előállítás, újraszervezés, invertálás stb.) is rendkívül költséges. Csak akkor válik gazdaságossá az adatbázis működtetése, ha különböző szolgáltatások egyaránt felhasználhatják azt. Ez nagyfokú normalizálást igényel. Új lehetőségeket rejt magában az öntanuló rendszerek alkalmazása: az ilyen rendszer az igények folyamatos figyelembevételével újraszervezi a file-okat, ezzel csökkentve a soron következő igények kielégítésének költségeit.

Kutatás

A nagyméretű integrált adatbázisok nemcsak hatékonyság és gazdaságosság szempontjából előnyösek, hanem sok olyan melléktermékük van, amely nagy jelentőségű lehet. Így például a nagyméretű adatbázisokon végzett statisztikai vizsgálatok segítséget nyújtanak tudománypolitikai feladatok megoldásához, folyóiratok értékeléséhez, szótárak összeállításához stb.

Más vizsgálatok pedig visszahatnak a file szerkesztés elveire. Az egyik legtöbbet ígérő vizsgálat kimutatta, hogy a *Chemical Titles* és CAIN (Cataloging and Indexing) adatbázisokban a 127 leggyakrabban előforduló címszó az összes előfordulásuk 44%-át tette ki. Ezekhez 8 bites kódokat rendelve 63%-os adat-kompresszió érhető el.

Másik lehetséges megoldást nyújt a tömegtárolók alkalmazása. A konferencián ismertettek egy 1012 bit nagyságú fotodigitális tároló használatát is.

A műszaki és gazdasági nehézségek leküzdésével kialakított nagyméretű adatbázis rendszerek azonban olyan pszichológiai, szociológiai, politikai és egyéb problémákat is felvetnek, amelyek megoldása további nagy erőfeszítéseket igényel.

/Journal of Chemical Information and Computer Sciences, 15. k. 1. sz. 1975. p. 3–5./

(Valkó Péter)

* * *

Környezetügyi adatrendszerek az Egyesült Államokban

Az Egyesült Államok Legfőbb Állami Számvevőszéke (General Accounting Office, GAO) a kongresszus megbízásából felülvizsgálta a környezetügyi adatok gyűjtésének és tárolásának helyzetét.

A vizsgálat kiterjedt az adatok fajtáira, az adatgyűjtés körére, az információátvitel módjaira, az információk hozzáférhetőségére a magán- és a közületi szektor számára, beleértve a kormányhivatalokat is; továbbá az adatgyűjtés koordinálásának módszereire és az adatoknak rendszerek közötti cseréjére.

8 minisztérium irányítása alá tartozó 320 adatközponttól és 10 önálló, független intézménytől érkezett válasz. Valamennyi gyűjt és tárol környezetügyi adatokat, többben az egész országra kiterjedően. Jelentős részük ugyanazon a részterületen, pl. 120 intézmény a levegővel kapcsolatosan gyűjt és tárol adatokat, ezeket a legtöbb rendszerben gépi adathordozókon, pl. lyukkártyákon és mágnesszalagokon tárolják.

Bár a szövetségi kormánynak nincs központi intézménye az adatok cseréjének és gyűjtésének összefogására, 167 intézmény jelezte, hogy rendelkezésre bocsát adatokat és információkat, 94 intézmény pedig kap adatokat más rendszerektől. 98 intézmény tagja olyan hálózatnak, amelyek az adatok cseréjére, illetve a gyűjtőtevékenység koordinálására alakultak. Egy esetben jelezték, hogy az adatok cseréje automatikusan, számítógépek között történik. Egyéb használatos eszközök, módszerek: mágnesszalag, kihelyezett terminál, papírmásolat, közvetlen kapcsolatok.

Az adatcserében információs központok felállításával kívánnak előrelépni. Terveznek pl. számítógépes master-indexet a vizekre vonatkozó adatforrásokról.

Más rendszerek adatainak beszerzésében és hasznosításában az alábbi nehézségeket sorolták föl az érintettek:

- az egyes rendszerekben használatos adatelemek és kódok nem kompatibilisek;
- az egyes rendszerekben alkalmazott berendezések és információt őrző nyelvek nem kompatibilisek;
- a kért adatokat gyakran késlekedve szolgáltatják;
- a más rendszerekből származó adatok pontosságát, megbízhatóságát nehéz fölmérni;
- az adatok újrakódolása költséges.

/Information Retrieval and Library Automation, 10. k. 9. sz. 1975. p. 5–6./

(Fazekas Zsuzsanna)

