A Chemical Abstracts adatbázis-kivonatainak szövegében végzett információkeresés hatásfoka

Bevezetés

A szerzők korábban (1976-ban) a *Chemical Abstracts (CA)* adatbázis két állományának – a *Chemical-Biological Activities (CBAC)* és a *Polymer Science and Technology (POST)* – mágnesszalagjaiban már végeztek keresésihatásfok-vizsgálatokat. A vizsgált rekordok a CA valamennyi adatelemét, így a kivonat szövegét is tartalmazták. A *Chemical Abstracts Service (CAS)* 1975 előtt csak ennek a két állománynak a kivonatait tette elérhetővé mágnesszalagon, 1975 óta már mind a 80 CA-szekció kivonatait mágnesszalagra rögzítik [1, 2].

Mások is vizsgálták már a kivonatban történő keresés hatásfokát a CA adatbázisban [3], de más adatbázisokban is [4]. Most a CA adatbázis különböző adatelemeiben végzett *szabad tárgyszavas* keresés hatásfokának vizsgálatát tűzték ki célul. A keresési stratégia megfogalmazásánál ugyanis a csak alkalmi online kereséseket végző kémiai képzettségű végfelhasználók és még inkább a nem vegyész szakemberek a CA esetén a szabad tárgyszavakra épülő keresést részesítik előnyben a kötött tárgyszavak, deszkriptorok használatával szemben. A CA esetén a kötött tárgyszavak alkalmazásának elsősorban a vegyületekre vonatkozó keresésben van jelentősége.

Az új vizsgálat eredménye azért is különösen érdekes, mert a CAS 1985 novemberében lehetővé tette a CAS ONLINE rendszeren a kivonatok szövegének online keresését, és mostanában sokan foglalkoznak az *American Chemical Society (ACS)* folyóiratokat tartalmazó adatbázisban végezhető *teljes szövegű* online keresés lehetőségeivel is [5, 6].

Módszer

A vizsgálat megkezdése előtt a CA 82-es kötetének (1975. január–június) CBAC és POST állományait olyan formátumra alakították, amely lehetővé tette többek között a jobb és bal oldali csonkolást, továbbá a *Chemical Substance Index* (vegyületmutató), valamint a *General Subject Index* (általános tárgymutató) valamennyi adatelemének keresését. Tíz keresőkérdést választottak ki véletlenszerűen, nyolcat a CBAC és kettőt a POST állományra *(5. táblázat)*. A *keresőprofilokban* (keresőstratégiákban) kizárólag *szabad tárgyszavakat* használtak, amelyek véletlenül egybeeshettek a CA indexkifejezéseivel. CAS Registry Numbereket (a vegyületeket azonosító kódszám) természetesen nem használtak.

A futtatásokat három adatmezőben végezték el *külön-külön:* a) a cím- és kulcsszómezőben (Keyword Phrases*) – *TK;* b) az indexmezőben (Chemical Substance és General Subject Index) – *I;* c) a kivonatmezőben (Abstract) – *A*. A relevanciát a teljes rekord tartalma alapján határozták meg, mind az egyes adatmezőkben, mind az összes adatmezőben végzett keresésre.

Példák

A "Benzodiazepinek az altatásban" témában végzett keresés jól illusztrálja az alkalmazott módszert és eredményeit *(1. ábra)*.

Az 1. és 2. keresőkifejezés keresőszavait az OR logikai operátorral, majd az 1. és 2. keresőkifejezés eredményét az AND logikai operátorral kapcsolták össze (így keletkezett a 3. keresőkifejezés). A keresés eredményét az *1. táblázat* tartalmazza.

1. táblázat

A "Benzodiazepinek az altatásban" téma keresésének eredménye[a]

CAN[b]	TK	I	A
82:000241	+		+
82:025830	+		
82:038611	+		
82:068285	+	+	+
82:080584	+ (F)	+ (F)	+ (F)
82:144982	+	+	+
82:038651		+ (F)	
82:093143		+	+
82:106282		+	
82:164862		+	+
82:011204			+
82:051497			+
82:051638			+ (F)
82:064536			+
82:093286			+ (F)
82:175185			+ (F)

[a] (+) = összes találat, (F) = nem releváns találat (zaj)
[b] CAN = CA hivatkozási szám (kötet- és referátumszám a nyomtatott CA-ban)

* Az egyes CA-füzetek végén található tárgymutató (nem kötött).

1. keresőkifejezés	:BENZODIAZEPIN:	:EPAM:
	:AZEPATE:	:EPAN:
	:DIAZEPOXIDE:	:OLAM:
2. keresőkifejezés	:ANESTHE:	:SADDLE BLOCK:
	:PREMEDIC:	:SADDLE-BLOCK:
	:PREOPERAT:	:LYTIC COCKTAIL:
	:OPERATION:	:LYTIC-COCKTAIL:
	:SURGER:	:HALOTHAN:
	:SURGIC:	:THIOPENT:
	:DENT:	:THIAMYLAL:
	:ODONT:	:METHOHEXITAL:
	:PAIN:	:METHITURAL:
	:NEUROLEPTANALG:	:FLURAN:
	:NEUROMUSCULAR BLOCK:	:NITROUS OXIDE:
		:N20:

3. keresőkifejezés = 1. keresőkifejezés AND 2. keresőkifejezés
: a csonkolás jele.

1. ábra Keresési stratégia
a "Benzodiazepinek az altatásban" témára

A *2. táblázat* összesíti a találatok számát, amelyeket bizonyos adatmezőcsoportokban, illetve a kivonatban végzett keresés eredményezett, megadva a keresés hatásfokát is.

2. táblázat

Az egyes adatmezőcsoportokra vonatkozó keresési hatásfok

	TK + I + A	TK + I	A	Kizárólag a kivonat (A) által behozott találatok
Releváns találatok	11	8	8	3
Nem releváns találatok	5	2	4	3
Összes találat	16	10	12	6
Pontosság (%)*	68,8	80,0	66,7	50,0
Teljesség (%)**	100,0	72,7	72,7	27,3

$$* \text{ pontosság} = \frac{\text{releváns találatok x 100}}{\text{összes találat}}$$

$$** \text{ teljesség} = \frac{\text{kihozott releváns találatok x 100}}{\text{összes releváns találat az adatbázisban}}$$

A TK + I + A oszlop az összes adatmezőben végzett keresés eredményét (releváns, nem releváns és összes találat) tartalmazza. A TK + I a cím-, a kulcsszó- és az indexmezőben végzett keresésre vonatkozik. Ezek egyébként a számos szolgáltatóközponton (Dialog, SDC, Data-Star, BRS stb.) elérhető CA SEARCH adatbázis kereshető adatelemei. Az *A* oszlop a kivonatban végzett keresés eredményét mutatja (átfedések lehetnek a két előző oszlopban kapott találatokkal).

A *2. táblázat* utolsó oszlopa azt mutatja, hogy a kivonatban való keresés behozott 3 olyan releváns találatot (az összes releváns hivatkozás 27,3%-a), amely a többi adatmezőben végzett kereséssel nem jött elő.

Az előző példában leírt módon vizsgálták a "klórtartalmú vegyületek LD50 értéke, illetve akut toxicitása" témát. A *2. ábra* mutatja a keresési stratégiát, a *3. táblázat* pedig a keresés eredményét.

1. keresőkifejezés	:CHLOR:	
2. keresőkifejezés	:LD50:	:TL50:
	:LD 50:	:TL 50:
	:DL50:	:LETHAL TOX:
	:DL 50:	:LETHAL DOS:
		:ACUTE TOX:

3. keresőkifejezés = 1. keresőkifejezés AND 2. keresőkifejezés

2. ábra Keresési stratégia
a "Klórtartalmú vegyületek LD50 értéke" témára

3. táblázat
A "Klórtartalmú vegyületek LD50 értéke" téma keresésének eredménye[a]

CAN	TK	I	A	CAN	TK	I	A
82:026809	+		+	82:081422			+
82:026834	+			82:081430			+
82:052359	+			82:081431			+
82:052366	+		+	82:081502			+ (F)
82:081391	+			82:092888			+
82:119681	+			82:093137			+ (F)
82:011275			+	82:093916			+
82:011949			+	82:093933			+
82:011957			+	82:093959			+ (F)
82:011958			+	82:094195			+
82:025653			+	82:106145			+ (F)
82:026048			+ (F)	82:106282			+ (F)
82:026706			+	82:106487			+
82:038624			+	82:106521			+
82:038740			+	82:107149			+
82:039241			+	82:107170			+
82:039601			+ (F)	82:107184			+
82:051346			+	82:107344			+
82:051586			+	82:119481			+
82:051673			+	82:119799			+
82:052302			+	82:120106			+
82:052349			+	82:132776			+
82:052377			+	82:132796			+
82:052582			+ (F)	82:133647			+
82:068177			+ (F)	82:133668			+
82:068250			+	82:133997			+
82:068948			+	82:147184			+
82:069004			+	82:149236			+ (F)
82:069173			+	82:149333			+
82:080343			+ (F)	82:149997			+
82:080352			+	82:164612			+
82:080592			+ (F)	82:164778			+
82:080660			+ (F)	82:165605			+
82:081390			+	82:165836			+ (F)
82:081421			+	82:165858			+

[a](+) = összes találat, (F) = nem releváns találatok (zaj)

A *4. táblázat* szemlélteti az egyes adatmezőcsoportokra vonatkozó keresési hatásfokot.

Az LD50 értékre vonatkozó példában a kizárólag kivonatban végzett keresés eredményezett 50-et az összes 56 releváns találatból (89,3%). A benzodiazepinre vonatkozó példában a kivonat önmagában 11 releváns találatból csak hármat hívott elő (27,3%). Az LD50 értékre vonatkozó példa világosan szemlélteti, hogy bizonyos esetekben a kivonatban végzett keresés jelentősen megnövelheti a releváns találatok számát és a teljességet.

Következtetések

Az *5. táblázat* foglalja össze a vizsgálatban szereplő mind a 10 keresőkérdésre vonatkozó eredményeket.

Látható, hogy a kivonatban való keresés átlagosan 50,2%-kal növelte meg a TK + I adatmezők által behozott releváns találatok számát.

Úgy tűnik, hogy a kivonat keresésének hasznossága függ a kérdés típusától és a rendelkezésre álló kötött tárgyszavak szelektivitásától. Az olyan kérdések esetén hasznos különösen a kivonat keresése, amelyekre ritkán vagy egyáltalán nem szerepelnek tárgyszavak az indexekben, mint például biológiai adatok és számszerű paraméterek, mint az "LD50". Olyan esetekben viszont, amikor a téma jól kereshető tárgyszavakkal is, a keresés kiterjesztése a kivonatra hátrányos lehet, nagymértékben csökkentheti a pontosságot.

4. táblázat

Az egyes adatmezőcsoportokra vonatkozó keresési hatásfok

	TK + I + A	TK + I	A	Kizárólag a kivonat (A) által behozott találatok
Releváns találatok	56	6	52	50
Nem releváns találatok	14	0	14	14
Összes találat	70	6	66	64
Pontosság (%)*	80,0	100,0	78,8	78,1
Teljesség (%)**	100,0	10,7	92,9	89,3

5. táblázat

Az egyes adatmezőkre vonatkozó találatok és a keresési eredmény hatékonyságának összehasonlítása 10 keresőkérdésre[a]

A kérdés tárgya	a TK + I + A keresésével	a TK + I keresésével	az A keresésével	az A keresésével válogatva
Festékek rögzítése (P)	45 (77)[b]	21 (27)[b]	39 (70)[b]	24 (50)[b]
Polisztirol homopolimer és korom (P)	10 (24)	5 (7)	9 (21)	5 (17)
Primidin típusú növekedési hormonok (C)	7 (12)	5 (10)	3 (3)	2 (2)
Terpének mint illatszerek (C)	19 (22)	10 (10)	16 (19)	9 (12)
Vakcinák (C)	47 (47)	34 (34)	32 (32)	13 (13)
Benzodiazepinek az altatásban (C)	11 (16)	8 (10)	8 (12)	3 (6)
Barbiturátok LD50 értéke vagy akut toxicitása (C)	2 (11)	0 (0)	2 (11)	2 (11)
Klórtartalmú vegyületek LD50 értéke vagy akut toxicitása (C)	56 (70)	6 (6)	52 (66)	50 (64)
Acetanilid típusú herbicidek (C)	32 (35)	24 (26)	25 (28)	8 (9)
Növények fotoszintézise (C)	50 (107)	26 (32)	39 (90)	24 (75)
Összesen	279 (421)	139 (162)	225 (352)	140 (259)
Pontosság (%)	66,3	85,8	63,9	54,1
Teljesség (%)	100,0	49,8	80,6	50,2

[a] (P) = POST; (C) = CBAC [b] zárójelben az összes (releváns + nem releváns) találat száma látható.

Irodalom

[1] BUNTROCK, R. E.: Searching Chemical Abstracts vs. CA Condensates. = Journal of Chemical Information and Computer Sciences, 15. köt. 3. sz. 1975. p. 174–176.

[2] BLAKE, J. E.–MATHIAS, V. J.–PATTON, J.: CA Selects – A Specialized Current Awareness Service. = Journal of Chemical Information and Computer Sciences, 18. köt. 4. sz. 1978. p. 187–190.
BLAKE, J. E.–EBE, T.: Abstract text searching for CA Selects. 2nd Chemical Congress of the North American Continent, 180th National Meeting of the North American Chemical Society, Las Vegas, Nevada, 1980.

[3] BARKER, F. H.–VEAL, D. C.–WYATT, B. K.: Comparative efficiency of searching titles, abstracts and index terms in a free-text database. = Journal of Documentation, 28. köt. 1. sz. 1972. p. 22–36.

[4] WAGERS, R.: Effective searching in database abstracts. = Online, 7. köt. 5. sz. 1983. p. 60–77.

[5] DURKIN, K.–EGELAND, J.–GARSON, L. R.–TERRANT, S. W.: An experiment to study the online use of a full-text primary journal database. = 4th International Online Information Meeting, 1980. London.

[6] COHEN, S. M.–SCHERMER, C. A.–GARSON, L. R.: Experimental program for online access to ACS primary documents. = Journal of Chemical Information and Computer Sciences, 20. köt. 4. sz. 1980. p. 247–252.

/HERZ, M.–KAINDL, H. K.–SALIB, A. A.–WARSZAWSKI, R.: Comparative efficiency of searching abstract text in the Chemical Abstracts Service Database. = Journal of Chemical Information and Computer Sciences, 25. köt. 2. sz. 1985. p. 111–114./

(Novák Teréz)