INFORMÁCIÓKERESŐ NYELV

Dokumentumelemzés és információkereső nyelv

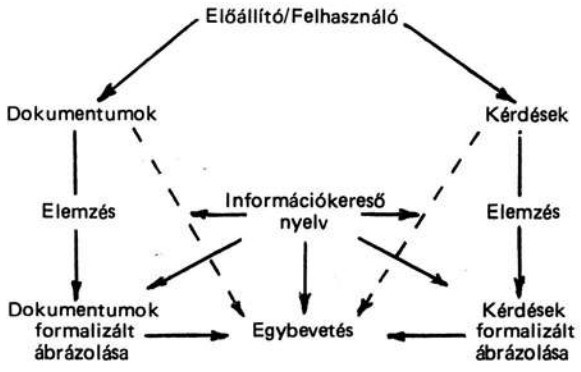
Az információnak az előállítótól a felhasználóhoz való eljuttatásában a legfontosabb állomás: megkeresni és kivonni a dokumentumokból a leglényegesebbet, s ezt a lényeget olyan formába önteni, hogy később minél gyorsabban, könnyebben tudjunk egy-egy információt kikeresni. Ennek alapvető eszköze az információkereső nyelv.

1. Az információ áramlása, az információkereső nyelvek szerepe

1.1 Az információ áramlása

Ha elfogadjuk, hogy a tájékoztatási funkció jelentése a szükséges dokumentumoknak a felhasználók rendelkezésére bocsátása, akkor megállapíthatjuk azt is, hogy a *dokumentációs tevékenység célja az, hogy kapcsolatot teremtsen a dokumentum és a felhasználó között (1. ábra).* Az információ előállítója hozza létre azokat az elsődleges dokumentumokat, amelyekben a felhasználó kérdéseire lehetséges válaszok találhatók. A felhasználó megfogalmazza kérdéseit, amelyekre a dokumentumok halmazából várja a feleletet. A kérdés és a dokumentumhalmaz egybevetésével próbálják kiszűrni a halmazból mindazt, ami releváns lehet a felhasználó kérdése szempontjából. Ebben a folyamatban a dokumentumok már nem eredeti formájukban, hanem sűrítve, tömörítve, formalizálva jelennek meg. A formalizálást *osztályozási rendszerek, tezauruszok, deszkriptorok (összefoglalva: információkereső nyelvek)* segítségével végzik.

Természetesen *az 1. ábrán bemutatott áramlási séma nem az egyetlen lehetséges út az információ előállítója és felhasználója között.* Az út többféleképpen rövidíthető meg. (Pl. ha a kutató, miután tanulmányozta a katalógust vagy indexet, közvetlenül a dokumentumok formalizált változatát használja és nem formalizálja kérdését. Előfordul az is, hogy a dokumentum tartalmát nem formalizálják stb.). A leghatékonyabb azonban mégis az *1. ábrán* bemutatott megoldás, azaz ha mind a dokumentumokat, mind a kérdéseket formalizálják. Így a keresés teljesen mechanikus műveletté válhat, számítógéppel vagy egyszerűbb eszközökkel is gyorsan, pontosan elvégezhető. Elsősorban ennek köszönhető a nagy rendszerek, pl. a CAN/SDI, SDC, LOCKHEED, PASCAL, MEDLARS, MEDLINE rohamos fejlődése.



1. ábra **Az információ áramlása**

Az ilyen rendszerekben az eredmény minőségét alapvetően a dokumentumok és a kérdések feldolgozásának módja határozza meg. Az információkereső nyelvek egységes használata jelentősen javítja a feldolgozás minőségét, és ezzel együtt a dokumentációs szolgáltatások minőségét is meghatározza.

1.2 Az információkereső nyelv fogalma

Az információkereső nyelv mesterséges nyelv, amelynek segítségével lehetővé válik a dokumentumok tartalmának és a kérdéseknek formalizálása, a felhasználók kérdéseire releváns válaszokat tartalmazó dokumentumok keresése. A nyelv egységekből, szimbólumokból áll, amelyek eredeti jelentésükön túl más jelentést hordoznak (például a Dewey-féle rendszerben *100* a filozófiai dokumentumokat jelöli, míg a tezauruszokban egy szó – betűkép – meghatározott fogalmat jelöl). *A szimbólumok mindig önkényesek,* általában nincs semmi összefüggés a betűkép és a jelentések között. A nagy osztályozási rendszerek használatánál ez közismert, de tudomásul kell vennünk, hogy a deszkriptorok kiválasztása a tezauruszoknál ugyanilyen önkényes még akkor is, ha az egyszerűség kedvéért a természetes nyelv szavait használják fel. Az információkereső nyelvek többségének van szintaxisa (szintaxis = a szimbólumok használatára vonatkozó szabályok összessége).

A mesterséges nyelv azt jelenti, hogy meghatározott cél érdekében tudatosan állították össze. Vannak olyan általános fogalmakból alkotott nyelvek, amelyekkel bármilyen szakterületről származó bármilyen dokumentumot vagy kérdést fel tudnak dolgozni. Másokat viszont egy-egy szűkebb szakterületen használnak (például az ágazati tezauruszokat).

Ez a cél határozza meg *az információkereső nyelv és a természetes nyelv közötti alapvető különbségeket.* Például: a mesterséges nyelv egységeinek, a szimbólumoknak egyértelműeknek kell lenniök; egy szimbólum csak

egyetlen fogalmat jelölhet; egy fogalom csak egyetlen szimbólummal fejezhető ki. Az információkereső nyelv szintaxisa is egyszerűsödik a célnak megfelelően. A dokumentumok és a kérdések formalizálása éppen ezen egyértelmű szimbólumok és egyszerűsödött szintaxis segítségével lehetséges. Ebből következik viszont az, hogy *egyszerű ábrázolhatóságuk miatt ezek számítógéppel is kezelhetők.*

2. A dokumentumelemzés folyamata

A dokumentumelemzés (a dokumentumok tartalmi feltárása) az információkereső rendszer működésének előfeltétele. Ezzel kapcsolatban azonban nyelvészeti és dokumentációs problémák merülnek fel.

2.1 Nyelvészeti problémák

Ha a természetes nyelvű szöveg tartalmát az információkereső nyelven akarjuk kifejezni, a szóval és a szöveggel kapcsolatban jelentkeznek nyelvészeti problémák.

A szó egymáshoz értelmesen kapcsolódó betűk sorozata (nem foglalkozunk itt a szigorúan vett morfológiai problémákkal). Két alkotóeleme van: a *signifiant (jelölő) és a signifié (jelölt).* A signifiant a betűk (hangok) együttese: önálló nyelvészeti egység, konkrét realitás, közvetlenül fizikailag felfogható. A signifié a fogalom, az amire gondolunk, amikor egy szót használunk: elvont, csak a beszélő vagy hallgató tudatában létezik. Pl. a birka szónál a b.i.r.k.a. betűk alkotják a signifiant-t, a signifié pedig: „legeltetett állat, amely gyapjút és fogyasztásra alkalmas húst ad". Egy signifiant-nak több signifié-je lehet és fordítva (a könyvtár szó pl. jelenthet egy épületet, intézményt, gyűjteményt stb.).

A signifiant kiválasztása önkényes. A dokumentációs munkában azonban ennek nincs jelentősége, mert a tartalmat leíró betűkép, szimbólum mint signifiant ugyancsak önkényes. A dokumentumelemzésnél a legfőbb probléma tehát éppen az, hogy azonosítani kell a csak a használók tudatában meglévő fogalom-egységeket ezekkel a szimbólumokkal, és el kell érni, hogy az indexelők és a felhasználók ugyanazokat a fogalmakat használják. Az információkereső nyelv kialakítása során állítják össze a fogalomjegyzéket úgy, hogy az ebben szereplő *minden egyes fogalomnak, signifié-nek egy és csakis egy signifiant feleljen meg.* Azokon a területeken, ahol ezeket a fogalmakat még nem tisztázták elég világosan és egyértelműen (pl. a társadalomtudományokban), a szókészlet összeállítása és ellenőrzése rendkívül nehéz.

A szöveggel kapcsolatban az az elsődleges követelmény, hogy a különbözőképpen megfogalmazott, de azonos tartalmú szövegeket azonos módon kell indexelni. Ez azért nehéz, mert az új információt, amelyet a szerző közölni akart, általában *szintaktikai, szemantikai kombinációkkal* fejezi ki és nem izolált szavakkal. Egy adott rendszerben tehát előre ismert kifejezésekkel kell feltenni olyan kérdéseket, hogy az eddig ismeretlen információt megkapják, azaz azt az információt, amelyet már nem tudnak kifejezni az ismert kifejezésekkel. Pl. figyeljük meg a négy következő mondat közötti szintaktikai és szemantikai különbségeket:

1. a rozsda rongálja a vasat;
2. az oxigén hat a vasra és így keletkezik a rozsda;
3. az oxigénnek a vasra gyakorolt hatása hozza létre a vasoxidot, amit rozsdának is hívnak;
4. a vas oxidációja hozza létre a vasoxidot.

Itt például a lényeges információ, az oxidációs folyamat csak egyetlen mondatban fejeződik ki egy szóval, a többiben szintaktikai és szemantikai szerkezetekkel.

A rosszul (nem egyértelműen) megalkotott szókombinációk ugyancsak sok problémát okozhatnak az elemzés során. A könyvtárigazgatásról tartott valamely szemináriumról szóló dokumentum három szóval indexelhető: könyvtár, szeminárium, igazgatás. Ha viszont ezeket szintaktikai kapcsolat nélkül teszik egymás mellé, mást és mást jelenthetnek. Pl.: szemináriumi könyvtár igazgatása; könyvtár a szemináriumok igazgatásáról stb. Tökéletes megoldást erre a problémára máig sem sikerült találniok a nyelvészeknek, de részleges megoldást adnak az indexelésnél bevezetett *kapcsolatjelölők, szerepjelölők.*

2.2 Dokumentációs problémák

A dokumentum tartalmi feldolgozása során jelentkező négy alapvető problémakör: *az elemzés mélysége, az ítélet (döntés), az egységesség, a felesleges párhuzamos feldolgozás.*

Minél pontosabban (mélyebben) akarnak indexelni egy dokumentumot, annál nagyobb annak a veszélye, hogy olyan részleteket is indexelnek, amelyekre a felhasználónak nincs szüksége, amelyek nem fontosak. *A dokumentumokat tehát tartalmuk szerint, de a felhasználók keresési szempontjainak figyelembevételével kell indexelni.*

Az indexelőnek *döntéseket* kell hoznia ahhoz, hogy a kommunikáció hatékonyságát biztosítsa, nemcsak a szöveg tartalmát, hanem a tartalom vélt (becsült) felhasználhatóságát is figyelembe véve.

A döntésnek a kritériumai viszont nincsenek egyértelműen meghatározva, *nem egységesek,* talán nem is lehetséges egyértelmű megfogalmazást adni. Éppen ezért az indexelők ugyanazt a dokumentumot általában eltérő deszkriptorokkal indexelik, és csak kb. 50%-ban használnak azonos deszkriptorokat.

Az egységesség hiánya vezetett *a centralizált indexelő szolgáltatások* kialakulásához. Ezzel egyidőben azonban egyre szaporodnak a kisebb jelentőségű és helyi vállalkozások is, és így ugyanazt a munkát több helyen végzik el, feleslegesen, párhuzamosan.

3. Az információkereső nyelv tipológiája

A fentiekben vázolt nehézségek kiküszöbölésére az információkereső nyelvek többféle típusát alakították ki. Ezeket itt csupán a koordináció, az ellenőrzés és a pontosság szempontjából tárgyaljuk.

3.1 A koordináció

E szempont szerint megkülönböztethetünk: prekoordinált és postkoordinált indexelő nyelvet.

A prekoordinált nyelvben a specifikus (mély, részletező) fogalmak tárgyszavait még a dokumentum feldolgozása előtt állítják össze, esetleg több kifejezésből. Általában ezt a típust használják a könyvtárakban. Ez az információkereső nyelv kevés deszkriptorral dolgozik és a keresett információt egyetlen keresés alapján adja meg.

A postkoordinált nyelvben a fogalmakat az adattárolás után, a keresés pillanatában koordinálják. Ez a nyelv technikailag szükségessé teszi kiegészítő adattárak, az ún. *invertált adattárak* használatát. Az ilyen típusú nyelv alkalmazásakor a dokumentumhoz több oldalról közelítenek. A keresés két lépésből áll: a megfelelő dokumentumok azonosítása a kereső fogalmak szerint, majd ezek postkoordinálásával a több fogalommal kifejezett kérdésnek megfelelő dokumentumok lokalizálása.

3.2 Az ellenőrzés

Információkereső nyelvek készülhetnek *ellenőrzött és szabad szójegyzékek* alapján. Mindkettőnek vannak előnyei és hátrányai. Az első nagyobb beruházást igényel, hosszabb előkészítést. A másodiknál a keresés könnyebb. A használt szókészlet ellenőrzésére azonban bizonyos fokon mindenképpen szükség van. Még a szabad szójegyzékeknél sem használhatnak ugyanis minden kifejezést, hanem kiválasztják közülük azt, amelyik a legjobban és a lehető legegyértelműbben fejezi ki az egyes fogalmakat.

3.3 A koordináció és az ellenőrzés összekapcsolódása

E két szempont összekapcsolásával, nagyon leegyszerűsítve, négy típusú információkereső nyelvet különböztethetünk meg *(1. táblázat)*.

Prekoordinált, szabad szókincsű nyelvek

Gyakorlatilag és lényegében *a periodikus indexek nyelve* tartozik ide. Az indexeket számítógéppel állítják össze. A nyelv egyes lexikai egységei nem mindig fogalmat fejeznek ki, hanem gyakran valamilyen aktuális problémával, tudásanyaggal kapcsolatos témát. A felhasználó ezek segítségével könnyen megtalálja az őt érdeklő témára vonatkozó információt, a szabad szókincs pedig lehetővé teszi, hogy gyorsan, rugalmasan kövessék a napi hírekben megjelenő fogalomkincset.

Postkoordinált, szabad szókincsű nyelvek

Ide tartoznak az ún. *uniterm* (egykifejezéses rendszerek) és bizonyos statisztikai módszereket alkalmazó nyelvek. Az uniterm rendszerek csak kis dokumentációs rendszerben használhatók. A statisztikai módszerek segítségével pedig a szövegben, kivonatban előforduló kifejezések gyakoriságát megállapítva alakítják ki a nyelvet.

Osztályozási rendszerek

Ezek a legismertebbek és a legáltalánosabban használt nyelvek. Ide tartoznak az *egyetemes és a szakosított osztályozási rendszerek, a hierarchikus és a fazettás rendszerek.* Valamennyi ellenőrzött szójegyzéket használ, bennük egy tárgyat csak egy adott jel jelölhet. Prekoordináltak, és minden egyes jel – a hierarchiában elfoglalt helyének megfelelően – inkább csak egy szakte-

1. táblázat

Az információkereső nyelvek négy típusa

Nyelv	Prekoordinált	Postkoordinált
	1. típus	2. típus
Szabad szókincsű	Indexek nyelve	Unitermek Statisztikai módszereket alkalmazó nyelvek
	3. típus	4. típus
Ellenőrzött szókincsű	Osztályozási rendszerek Tárgyszórendszerek	Tezauruszok

rületre, témára utal, mint egy fogalomra. Hátrányuk általában az, hogy információkereséshez túl általánosak.

Tezauruszok

A tezauruszok az ellenőrzött szókincsű rendszerek legfőbb előnyét (pontosság) egyesítik a postkoordináció (rugalmasság) előnyeivel. Minden deszkriptort önmagában – a többire való hivatkozás nélkül – is lehet értelmezni és használni. A hierarchikus utalások is kevésbé kényszerítő erejűek, mint az osztályozási rendszereknél. Egyre világosabbá válnak viszont a tezauruszok alkalmazásának hátrányai. Így feltétlenül meg kell oldani az egyes nagyon speciális tezauruszok kompatibilissé tételének problémáját.

3.4 Az elemzés mélysége

Azt, hogy egy intézmény milyen információkereső nyelvet használ, elsősorban az dönti el, milyen mélységű elemzésre van szüksége. Ezzel kapcsolatban vizsgálandó az is, hogy egy információkereső nyelv mennyit tud átadni az eredeti dokumentum információiból.

Egy dokumentum tartalmát különböző módokon, például osztályozási jellel, kulcsszavakkal stb. lehet megadni. Sok esetben ezeknek ki kell egészíteniök egymást. A dokumentációs feldolgozás szempontjából ezért megkülönböztetik *a tudományterületet, a témát, a fogalmat és a magyarázatokat.*

A tudományterület olyan témák összessége, amelyeket kutatók egy csoportja alkalmasnak tart arra, hogy közös kutatás tárgya legyen. Ez lehet egy tudományág, de lehet interdiszciplináris jellegű is, széles vagy szűkebb terület.

A téma a Robert enciklopédia szerint „az, amiről gondolkodunk... a mindennapi szóhasználatban gyakran azonos a kutatás tárgyával..." Dokumentációs szempontból azonban a témát mindig a szerző határozza meg.

A fogalom mindig egy tudományág szakembereinek egyetértését feltételezi, mert a tudományos kutatás lényege a közös fogalmak meghatározása, elfogadása és az ezekre vonatkozó, az adott tudományterületeken jelentkező törvények tanulmányozása. Dokumentációs szempontból pedig az lényeges, hogy egy fogalmat – éppen azért, mert közmegegyezés tárgya – és a fogalommal kapcsolatos információkat minél könnyebben lehessen keresni.

A magyarázat azt jelenti, hogy a szerző új tapasztalatot, jelenséget ír le, új elméleteket állít fel, időbeli és okozati összefüggést fedez fel a jelenségek között, és ezeket leírja, de ehhez már nem a saját szakterületének szókincsét használja fel. A dokumentalista munkája ezen a ponton nagyon nehézzé válik, mert új, eddig ismeretlen információkat kell kifejeznie a rendelkezésére álló információkereső nyelvben meglévő, ismert kifejezésekkel.

A dokumentumok tartalmát tehát általánosságban a fentiekben részletezett négy információ-kategóriával ábrázolhatjuk: tudományterület, téma, fogalmak és magyarázatok.

A legáltalánosabban elterjedt ábrázolási szint az *osztályozási rendszerek* szintje. Az ETO vagy a fazettás rendszerek mélyebb feldolgozást tesznek lehetővé.

A dokumentációban dolgozó indexelő szakemberek azonban nem elégszenek meg azzal, hogy a dokumentumot csak a szerző szempontjából közelítik meg. A szövegben közölt specifikusabb információkat is igyekeznek megadni, olyan fogalmakat használva, amelyeket a felhasználók kereshetnek. Ennek az ábrázolási szintnek megfelelő eszköz *a tezaurusz* és az azokban foglalt deszkriptorok. *Az unitermek* használata is megfelel ennek a szintnek, de terminológiájuk még nem eléggé egységes.

A magyarázatok szintjén nem alkalmazható az itt bemutatott információkereső nyelvek közül egy sem. A legjobb eszköz itt a természetes nyelv. Egy kivonat pl. – a magyarázat szintjén – az összes fontos információt megadhatja. Az informatív vagy indikatív kivonatok

2. táblázat

Az információkereső nyelvek tipológiája

Az elemzés szintje	Koordináció	Ellenőrzés	Információkereső nyelv
Sekélyebb			Osztályozási rendszerek LC, Dewey, ETO stb.
Tudományterület	Prekoordinált	Ellenőrzött szókincs	Tárgyszavak
Téma			Tezauruszok
Fogalom	Postkoordinált		Unitermek
Magyarázat		Szabad szókincs	Indikatív kivonatok
Mélyebb			Informatív kivonatok

gyakran egyszerű annotációk vagy címmagyarázatok, egy-egy cikk lényegét, célját mégis jól kifejezik.

Ha elfogadjuk azt, hogy a különböző szinteken átadott információ nem azonos természetű, érthető, hogy az ideális megoldást akkor találják meg a dokumentációs intézmények, ha a kívánt információnak legjobban megfelelő feldolgozási szintet (információkereső nyelvet) választják ki.

A *2. táblázat* azt mutatja, hogyan lehet kiválasztani a szükségleteknek legjobban megfelelő információkereső nyelvet. A táblázat természetesen nagyon leegyszerűsíti ezeknek az információkereső nyelveknek a jellemzőit. Egyetlen nyelv sem csak prekoordinált vagy csak postkoordinált, szabad vagy ellenőrzött szókincsű.

/COURRIER, Y.: Analyse et langage documentaires = Documentaliste, 13. köt. 5–6. sz. 1976. p. 178–189./

(Ferch Magda)

A rubrikátor mint információkereső nyelv

Bár a rubrikátor az osztályozás és információkeresés egyik elterjedt eszköze, információkereső nyelvként való alkalmazásának lehetőségét eddig nem elemezték.

A cikk célja a rubrikátor ilyen szempontból való elemzése, ami hasznos lehet mind a rubrikátort készítők, mind pedig a felhasználók számára.

Az elemzés kiindulópontja, az információkereső nyelv kritériumainak *A. I. Csernüj* szerinti meghatározása. Csernüj szerint *az információkereső nyelv a dokumentumok fő témakörének vagy formai jellemzőinek leírására alkotott mesterséges nyelv,* amely meghatározott dokumentumoknak valamely dokumentumhalmazból való kikeresésére és/vagy információs kérdések tartalmi kifejezésére szolgál.

Az alábbi gondolatmenet bizonyítja, hogy a rubrikátor megfelel az információkereső nyelv kritériumainak.

A rubrikátor lexikai, grammatikai felépítése lehetővé teszi bármilyen szöveg (dokumentum) fő témakörének kifejezését, pl:

17.02.02 Molekulák és kémiai kapcsolatok felépítésének elmélete

Minden rubrika két részből áll: *az indexből (számkódból) és a hozzá tartozó fogalmi magyarázatból.*

A fogalmi magyarázat a természetes nyelv szavait tartalmazza, de az általánosan használt rövidítések, valamint a kémiai, matematikai stb. képletek, jelek is használhatók. A fogalmi magyarázatokhoz tartoznak még a szükséges utalások, hivatkozások, keresztutalások is.

A rubrikátor grammatikáját a rubrikák közötti kapcsolatok alkotják. A rubrikák között a hierarchikus rendszerekre jellemző *nem–faj* kapcsolatok vannak, amelyek révén az általánosabb fogalmakból a részletek felé haladhatunk *(1. ábra).*

1. ábra

Nyilvánvaló, hogy a rubrikátor mondattani felépítését a rubrikák rubrikátorbeli sorrendje határozza meg. Az alárendeltségi viszony az indexek felépítésében is tükröződik. Esetünkben az alap-rubrika indexe *két számból* (17 Kémia) áll, minden további (alárendelt) szint újabb *három jellel* – ponttal és két számjeggyel – bővül.

Természetesen egy általános tematikájú rubrikátor nem képes a különböző ágazatok közötti kapcsolatokat kizárólag alárendelt viszonyításokkal kifejezni, ezért mellérendelésre utaló jelek (lásd, lásd még) alkalmazása is szükséges pl.

17.02.05 Fotokémia
A fizikában lásd 14.10.05 alatt.

Az ideális információkereső nyelv másik követelménye, hogy *kifejezései ne legyenek két- vagy többértelműek,* mert a keresés ebben az esetben két vagy több (felesleges) rubrikához vezetne. Erre a követelményre a rubrikátor összeállításánál nagyon kell ügyelni.

Az információkereső nyelvvel szemben támasztott további követelmény, hogy *elemeinek nem szabad szerzőre vagy esetleges címzettre utalni,* azaz az ideális információkereső nyelvnek teljesen „elfogulatlannak" kell lennie.

A rubrikátor ennek a követelménynek is megfelel. Nyelve prekoordinált, mivel a rubrikákat az alkalmazást megelőzően, az információs anyagok természetes áramlása alapján állítják össze, s így objektív voltához nem fér kétség.

Végül követelmény még a *gépi keresés lehetősége,* amely szerint az információs nyelvnek alkalmasnak kell lennie a tárolt dokumentumok keresőképének és a keresett dokumentumok deszkriptorainak algoritmikus összevetésére.

A rubrikák indexeit egy adott dokumentum formalizált tartalmi kifejezésének, keresőképének is tekinthetjük. Az információs igényeknek a rubrikátor nyelvén történő leírása a kérdések formalizált kifejezése. Az automatizált keresés során a számítógép összeveti a kérdés deszkriptorait a rubrikák indexeivel és kiadja a