syi:

A nép szavai

Címkék és könyvtárak – a szemantikus web ígérete és valósága

A szematikus web fogalmát többféle módon is értelmezhetjük. Akárhogy is tesszük azonban, ez nem igazán befolyásolhatja azt az ítéletünket, amelyet az írásunk alcímével kapcsolatban fogalmazhatunk meg.[1] A szemantikus web ugyanis – véleményünk szerint – ma még sokkal inkább csak ígéret, mint valóság, és egyelőre nem is nagyon látszik, mikorra és hogyan leszünk képesek a szemantikus web program céljait megvalósítani. Ha tehát tartalmasat kívánunk szólni, akkor el kell tudnunk lépni az alcím sugallta témától.

A Szemantikus Web Kezdeményezés a W3C konzorcium programja. Személyesen Tim Berners-Lee nevéhez kapcsolódik, aki azt mondta az ezredfordulón, hogy amíg a web első szakasza arról szólt, hogy a számítógépeket megtanítottuk a szöveget olvasni, addig a következő évtizedben az a feladatunk, hogy megtanítsuk a gépeket arra, hogy képesek legyenek a szövegek értésére, értelmezésére is.[2] A cél érthető, fontos, nemes, ambíciózus, de a megvalósításától egyelőre fényévekre vagyunk. Az írásunkban éppen ezért nem a miértekről, nem a nehézségekről értekeznénk, mert a téma kifejtése elvinne minket a nyelvtechnológia, a szemantika, az ontológia irányába, amivel – érzésünk szerint – eltávolodnánk a konferencia fő témájától, de legalábis a potenciális közönségünk érdeklődésétől. Érdekesnek és fontosnak tartunk viszont egy olyan kérdést körbejárni, amely szemantikai problémák mentén kapcsolja össze azt a két jelenséget, amelyre a webről szóló közbeszédben web 2.0 és web 3.0 fogalmakkal szoktak hivatkozni.[3]

Az a meglepő helyzet ugyanis, hogy a szemantikus web kezdeményezés céljait jelölik manapság a web3 terminussal, míg a web történetében időben később elterjedő közösségi tartalomszolgáltatás jelenségegyüttesére alkalmazzák a web2 kifejezést.[4] Noha nem igazán tartjuk jónak ezt a két terminust, lehetne jobbakat használni helyettük (vannak is ilyenek, például a ‚peer production' fogalmával sokkal több jelenség sokkal pontosabban megragadható[5]), mégis megtartjuk ezt a fogalomkettőst a továbbiakban, hiszen napjainkban nagyon széles körben használatban vannak (főleg persze a web2 kifejezés). A megvizsgálandó kérdést pedig úgy fogalmazzuk meg, hogy vajon milyen kapcsolatot képzelhetünk el a web2 és web3 jelenségek közé, vagy másként fogalmazva és a gondolatmenetünknek némi provokatív élt adva: lesznek-e, s ha igen, mire lesznek jók a könyvtárosok a jövőben. Úgy ítéljük meg ugyanis, hogy a web2-es jelenségek – igen jelentős részben, bár koránt sem teljes mértékben – fölöslegessé teszik a könyvtárosok (professzionális archivátorok) munkáját, miközben a digitális archívumokban egyre nagyobb szemantikai vákum keletkezik, s ebből fakadóan egyre nagyobb szükséglet támad a hiányzó szemantikai információk pótlására, amit viszont könyvtárosokkal (professzionális archivátorokkal), illetve gépi szemantikával, gépi tanulással építhetünk be a rendszerbe.

Induljunk ki az információ, tartalom, metainformáció, dokumentum, archívum és keresés fogalmakból felépíthető fogalmi modellből. Ha az *információt tárolni* akarjuk, akkor a szóbanforgó *tartalmat* valamilyen hordozóra rögzítve *dokumentumot* (könyvet, képet, videót, hanglemezt stb.) hozunk létre. A tárolási tevékenység egyik kiemelt célja nyilván a tartalom későbbi befogadásának biztosítása, s ezt a célt a dokumentumok *archívumba* rendezésével érhetjük el a leghatékonyabban. A tartalom befogadhatóságához azonban meg kell tudnunk találni az archívumon belül a keresett dokumentumot, és e célból kiegészítő információt, *metaadatot* kell a dokumentumokhoz rendelnünk. A metainformáció elsődleges célja a dokumentum visszakereshetőségének a biztosítása az archívumon belül, s ebből fakadóan az archívum egyik legfontosabb tevékenysége a *keresés* (illetve a kereséstámogatás). Amióta archívumokat építünk magunknak (ideértve a könyvtáraktól kezdve a levéltárakon, filmtárakon át a vállalati dokumentumtárakat vagy a személyes könyv-, dvd- vagy cdgyűjteményeket), mindig és minden helyzetben alkalmazni lehetett ezt az egyszerű modellt.[6]

A tárolás és keresés gyakorlati problémáira az eddigi legkiérleltebb választ az évszázados könyvtári hagyományban találhatjuk meg. Nem véletlen, hogy a web kezdetétől fogva sokak számára tűnt

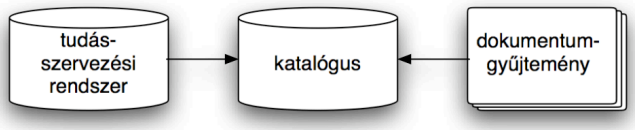
[1] A konferencián elhangzott előadás címét jelen írásmű alcímévé tettük, mert a tanulmányunk tartalmához jobban illeszkedő címet akartunk választani.
[2] [Berners-Lee et al. 2000]
[3] A web 2.0 és web 3.0 kifejezések helyett a továbbiakban a web2 és web3 rövidebb alakokat használom.
[4] A web 2.0 terminus megalkotását Tim O'Reilly-nak tulajdonítják [O'Reilly 2005].
[5] A peer production fogalma alá sorolható jelenségekről, a fogalom meghatározásáról bővebben: [syi 2007]
[6] A keresés modellezéséről bővebben: [syi 2007]

érdemesnek hasznosítani ezen értékes tudást és tapasztalatot. Az ilyen kezdeményezések azonban elég kevés sikerrel jártak mindeddig, aminek meg kellene találnunk a magyarázatát.
A továbblépés előtt érdemes még a modellünkön egy szempontból tovább finomítani, s legalább azt a megkülönböztetést átvenni, amit a könyvtári világ kiformált azzal, hogy a könyvek metaadataival kapcsolatos *formai* és *tartalmi feltáró* tevékenységeket elválasztotta egymástól.[7] A dokumentumok *formai metaadatait* csak azáltal tudjuk megfelelően kezelni, ha pontosan és egyértelműen minősítjük az egyes metaadatelemeket, ami egyenes arányban van a formai metadatrendszer strukturáltsági fokával (tehát bonyolultságával). Ahány formai adatelemet kezelni akarunk, annyiféle entitást és majdnem ugyanannyi relációtípust kell definiálnunk, hogy aztán azok konkrét értékeit a dokumentumokhoz rendelhessük. Más a helyzet a tartalmi leíró tevékenységgel. Többféle lehetőség van a *tartalmi metaadatok* könyvekhez rendelésére. A legelterjedtebb megoldás az, hogy a dokumentumok tartalmát valamilyen előre kidolgozott készletből válogatott kifejezésekkel, kulcsszavakkal, tárgyszavakkal jellemezzük. Amikor ezt tesszük, akkor a szóbanforgó terminuskészletet már érdemes önálló elemként elkülönítenünk az archívumi modellünkben, aminek megnevezésére használhatjuk a nemzetközileg elfogadott kifejezést: *tudásszervezési rendszer* (knowledge organization system, KOS).
A tudásszervezési rendszerek (KOS-ok) a tartalmi feltáró munka során a dokumentumokhoz rendelhető tárgyszavakból és az ezek között tételezett relációkból állnak. Matematikai értelemben azt mondhatjuk, hogy a KOS a tárgyszavak *tartóhalmazán* értelmezett *struktúra*, amelyből természetesen lehetséges többféle is annak megfelelően, hogy milyen *relációkat* engedünk meg felvenni a tárgyszavak között. A struktrúra az alábbi formában írható fel:

$KOS = <D, R_1, R_2, \ldots, R_n>$, ahol

D a tudásszervezési rendszer tárgyszavaiból álló tartóhalmaz
R_i az elemeken (tárgyszavakon) értelmezhető reláció

Felmerülhet a kérdés, miért is van szükségünk a relációk értelmezésére. Legalább két fontos funkcióra már előzetesen rámutathatunk, mégha ezeket csak a későbbiekben tudjuk majd bővebben kifejteni. A tárgyszó-hozzárendelés célja és értelme a dokumentumok minél pontosabb tartalmi leírása. A természetes nyelvünk rugalmasságának „ára“ a szavaink, kifejezéseink többértelműsége. A pontos tartalmi leíráshoz arra van szükségünk, hogy a nyelv, egyébként sokértelmű szavait, kifejezéseit egyértelmű jelentés mellett tudjuk a dokumentumokhoz rendelni. Ez az egyértelműsítés azáltal teremthető meg (legalábbis a hagyományos archívumi gyakorlatban), hogy a tárgyszavaknak megmutatjuk a fogalmi környezetét (vagyis azt, hogy más fogalmakkal milyen kapcsolatban áll). Ezért a tárgyszavazást végző személyeknek a tárgyszókészletet annak teljes struktúrájával együtt kell látniuk (a tárgyszavak közti relációk abban is segítséget adhatnak egyébként, hogy támogatják a rendszeren belüli navigációt a tárgyszókeresési munka során). A relációk másik fontos „funkciója“ pedig az lehet, hogy segítségükkel alternatív tárgyszavakat lehet esetleg megtalálni, illetve logikai következtetéseket lehet végrehajtani (ha ez szükséges).
Addig jutottunk tehát, hogy megállapítsuk: a metaadat-hozzárendelési munkát – a szemantikai egyértelműsítés végett – kétféle módon is támogatta a könyvtári gyakorlat. Egyrészt a formai adatelemekre egy sokdimenziós (sok relációból álló), nem túl összefüggő, minden dimenzióban nagyon lapos, de pontosan rögzített struktúrát teremtett, másrészt a tartalmi leírás egyértelműsítésére a tudásszervezési rendszerek elemeit vette használatba. A továbbiakban az utóbbi mozzanatra fogunk koncentrálni.
A tartalmi leíró tevékenység során a dokumentumokat a tudásszervezési rendszer elemeivel jellemezzük úgy, hogy összekötjük egymással az éppen elemezett dokumentumot a KOS-rendszer kiválaszott elemeivel. Ha az összekapcsolás tényét rögzítő adatot a többi információtól elkülönítve kezeljük, akkor létrehozzuk (fenntartjuk) a katalógus „intézményét“. Azt mondhatjuk tehát, hogy a tartalmi leíró tevékenység a katalógus céduláinak, rekordjainak írását jelenti. Az archívum általános modelljében három nagyobb információs blokk van, amelyeket az alábbi ábrával szemléltethetjük:



1. ábra: információblokkok az archívumban

[7] A dolgokat kissé leegyszerűsítve azt mondhatjuk, hogy a formai leíró elemeket megadhatjuk akkor is, ha nem ismerjük a dokumentum tartalmát, ellenben a tartalmi metaadatokat csak a tartalom ismeretében lehet megállapítani.

A vázolt fogalmi modellünkre támaszkodva most már nekiláthatunk mondandónk kifejtésébe. A történetet a web megjelenésével kell kezdenünk. A web egyik forradalmi újdonságát (és erejét) a *szabadszavas keresés* megjelenése adta, amelynek sikere sok embert annak kimondására sarkallt, hogy talán nincs is szükség másfajta keresési lehetőségre. Bár a szabadszavas keresés (főleg a relevanciakezelés különböző megoldásaival együtt) valóban nagyon jó eredményekre képes, korábban sosem volt lehetőségeket kínál számunkra, tudnunk kell, hogy nagyon komoly és kiküszöbölhetetlen hiányosságai vannak ennek a technológiának. A szabadszavas keresés legfontosabb problémája nyilván az, hogy csak szöveges dokumentumok esetében működőképes, vagyis az audiovizuális dokumentumok esetében nem használható.[8] De még „tiszta“ szöveg esetében is komoly gondok adódhatnak vele. Mivel a szabadszavas keresőkben nem tudjuk minősíteni a keresett kulcsszavakat, ezért nincs mód a metaadatok minősített keresésére (a szöveg strukturálatlanságával nem tud megbirkózni ez a technológia, emiatt a metaadat mint olyan kezelésére egyáltalán nem képes).

> A szabadszavas keresés során nem lehetséges a szerző vagy a dokumentum címe szerint keresnünk. Ha a dokumentumban „valódi“ címként szerepel egy karaktersorozat, míg egy másikban csak hivatkoznak ugyanerre a címre, akkor mindkettőt visszakapjuk – függetlenül a kifejezés „metainformációs státusától“.

Mondhatnánk persze azt, hogy a tartalmi feltárás területén viszont hatalmas előnyként jelentkezik az a lehetőség, hogy nem kell elvégezni a tárgyszavazás fáradságos munkáját, mivel az automatikus gépi indexelés megcsinálja azt az ember helyett. Ez kétségtelen tény, és ez valóban komoly előnyt biztosít, de azért ennek a megoldásnak is vannak hátulütői. A gépi indexelés ugyanis nem tud mit kezdeni a nyelv többértelműségével, amit viszont az ember a feltáró munka során kezelni tud. Nézzünk meg ezekből néhány példát!
A gépi keresés nem tudja elkülöníteni a *homonimikus* jelentéseket, akár a köznévi, akár a tulajdonnévi alakok közötti homonimákról van szó:

> macska (állat) – vasmacska (eszköz)
>
> Cica (‘Révész Cica József klarinétos’) – cica (macska)

A keresést támogatásának egyik fontos eszköze a szavak, kifejezések közötti *szinonimitás*, amellyel az automatikus indexelés semmit sem tud kezdeni. A számítógép számára nem nyilvánvaló (nem ismert) az alábbi három terminus (durván) azonos jelentése:

> macska - cicus - cica

Nem kell most elmerülnünk abban a vitában, amely a szinonima és a *poliszéma* elválasztásáról zajlik a nyelvészet területén, itt elég csak jeleznünk, hogy szavak bizonyos jelentései között gyakran nincs teljes átfedés (vagyis nem beszélhetünk szinonimitásról), mégis szoros kapcsolatot állapíthatunk meg közöttük. Az alábbi példa nem lehet szinonima, hiszen állatról és emberről van szó, mégis érezzük a két jelentés kapcsolódását egymáshoz (amit a gép már nem tud megragadni).

> cica (macska) – cica (jó nő)

Aztán a nyelv „továbblép“ még egyet, amikor a pozitív töltetű jelentést átértelmezi, s ironikus módon, ellenkező értelemben használja a terminust.

> cica (jó nő) – fitneszcica (kövér, csúnya nő)

Nehezebb észrevenni, ami a számítógép számára sokkal nyilvánvalóbb, hogy milyen sok olyan *akronimát* (betűszót) képzünk, amelyek a köznyelv valamely szavával teljesen megegyeznek.

> CICA (Criminal Injuries Compensation Authority)
>
> CICA (Confederation Internationale du Credit Agricole) – cica (macska)

A keresés során talán nem annyira fontos, de azért említésre érdemes az a képességünk (amivel a gépek egyelőre nem rendelkeznek), hogy azonnal felismerjük bizonyos szavak, kifejezések ellentétes jelentését, vagyis kezelni tudjuk az *antonima* nyelvi jelenségét.

> kutya – macska

A nyelv rugalmasságának egyik legfontosabb eszköze a *metonima*, amellyel sokféle értelemben kiterjeszthetjük egy szó jelentését. Az alábbi két példában az ‚iskola‘ két jelentése eltérő módon lenne

[8] Audiovizuális dokumentumon nem természetes nyelvi alapú, képi és/vagy hangi információkat tartalmazó dokumentumot értünk.

leírható egy formális ontológiában, noha érezzük, hogy ezek a jelentések szoros kapcsolatban vannak egymással.

Három szintes iskolába járunk. (iskola mint épület)

Nyolcosztályos iskolába járunk. (iskola mint intézmény, testület)

Tovább bonyolódik a helyzet, amikor a hálózat többnyelvűségét is figyelembe kell vennünk, s a fenti többértelműségek a különböző nyelvek között is létrejöhetnek. Előfordulhat „többnyelvű szinonimitás“, amikor különböző nyelveken fejezzük ki ugyanazt a jelentést, tartalmat:

macska (magyar)

cat (angol)

Katze (német)

A másik irányból tekintve értelmezhetünk „többnyelvű homonimitást“ is, amikor ugyanaz a szóalak két nyelven teljesen más jelentéssel bír.

cica (magyar)

ciça (portugál)

A fenti példák mind azt mutatták, hogy egyetlen karaktersorozat mennyiféle jelentéssel rendelkezhet, amit a számítógépek – egyelőre még – nem, az emberek viszont képesek megfelelő módon, vagyis a szemantikus gazdagságuk teljében értelmezni. A ‚cica‘ karaktersorozatnak legalább az alábbi különböző értelmezéseit adhatjuk meg (a felsorolás nyilván nem teljes):

magyar: cica (macska)

angol: CICA (tulajdonnevek)

magyar: Hilton-cica,cica-mica (szép nő)

magyar: fitneszcica (csúnya nő)

magyar: papírcica (hajtogatott cica)

magyar: porcica (kosz)

A számítógép mindezekről semmit sem tud, a ‚cica‘ terminust nem tudja sehogysem értelmezni, nyugodtan mondhatjuk azt, hogy teljes *szemantikus vakságban* szenved – ma még. A Szemantikus Web Kezdeményezés célja és értelme pont az, hogy a keresőmotorok számára szemantikus képességeket tudjunk kifejleszteni.
A szemantikai vakság mellett azonban van még egy másik, igen komoly problémája a webes keresőmotoroknak. A szabadszavas keresés ugyanis *teljes-szöveges keresést* is jelent egyben (fulltext search), amit abból a szempontból előnyösnek és kívánatosnak minősíthetünk, hogy ezáltal a szöveges dokumentumok teljes szókészlete kereshetővé és elérhetővé válik (ami különösen ritkább szavak esetében lehet nagyon hatékony), viszont a *relevanciakezelés* területén új nehézségek jelentkeznek (pontosabb lenne azt állítani, hogy a relevancia egész problémaköre teljesen új megvilágításba kerül). Amikor a szöveges dokumentum összes szavát leindexelik a keresőmotorok, és az invertált index alapján a szövegben előforduló bármely szó alapján képesek visszaadni a dokumentumot magát, akkor rögtön felmerül a kérdés, hogy az adott keresőszó vajon jól jellemzi-e az adott dokumentumot, vagyis kellően releváns-e a szó és a dokumentum kapcsolata. A válasz nyilvánvalóan csak az lehet, hogy minden szó nem jellemezheti egyforma erővel a dokumentumot, tehát szükség lenne egy olyan módszerre, amely a szavakat tartalmazó dokumentumokhoz valamilyen relevanciaértéket rendel, ami alapján a fontosabbnak minősített dokumentumokat előbbre lehet rangsorolni a találati listákban. Az első generációs keresőgépek a relavanciaértékeket a dokumentumban magában keresték. Ebben persze önkéntelenül is követték azt a hagyományos világból származó gyakorlatot, amely a dokumentumokat tárgyszavakkal leírhatónak gondolta. Ez az évszázados könyvtári hagyomány azt sugallta, hogy a teljes-szöveges indexállományból ki lehet választani néhány (vagyis kevés) releváns tárgyszót a dokumentum tartalmi leírására. A kérdés csak az volt, hogy amit a könyvtárosok jól meg tudtak oldani, ti. a kevés releváns tárgyszót kiválasztani, azt vajon hogyan lehet megtanítani a számítógépnek. A keresőmotorok fejlesztői az első időszakban (a web kezdetén) olyan szempontokra próbáltak meg figyelni, mint:

hányszor fordul elő a szó a dokumentumban (ha többször, akkor „többet ér“)

hol szerepel a szó a dokumentumban (ha az elején, akkor „többet ér“)

szerepel-e a szó a dokumentum címében, alcímében (ha igen, akkor „többet ér")

Ezek a kezdeti próbálkozások azonban kevéssé (vagy egyáltalán nem) voltak hatásosak. A Google volt az első a keresőmotorok világában, amely valóban működőképes relevanciakezelést valósított meg. A Google megoldása, PageRank azonban teljesen más szempontra figyelt, mint az elődei. A dokumentumok relevanciaértékenek számításakor ugyanis nem a dokumentumok tartalmát (a benne szereplő szavakat) vette figyelembe, hanem a dokumentumokban elhelyezett linkek (más oldalakra mutató utalások) tényét, számát, súlyát. A weboldalak készítőinek szubjektív ítéleteit lehetett ezáltal összegyűjteni és aggregálni valamiféle közösségi fontossági mutatószámmá. A relevanciakezelés nem jelent egyebet, mint valamilyen módon kifejezni, hogy adott dokumentum, adott kontextusban fontos egy személy vagy egy közösség számára.

A könyvtári világ addig más technikát alkamazott a relevanciakezelés problámájára. Amikor a könyvtárosok a dokumentumok tartalmi leírását végezték, akkor a rendelkezésükre álló, elméletileg összes lehetséges leíró tárgyszó közül kiválasztották az általuk legfontosabbnak tartottakat, és ezeket hozzárendelték a dokumentumokhoz. A leíró tárgyszavaknak ez a kiválasztása, szűrése egy szempontból nagyon hasonlított a keresőmotorok teljes-szöveges indexeléséhez. A könyvtárosok az előre rögzített tudásszervezési rendszer elemeiből válogatták ki a legfontosabbnak tartott leíró elemeket, vagyis számukra ugyanúgy rendelkezésre állt egy előzetes szóhalmaz, amiből aztán választaniuk kellett, mint ahogy a keresőmotorok is minden egyes dokumentumról felállították a dokumentum összes szavából álló szóhalmazt, és ennek elemeihez tudták hasonlítani a későbbi felhasználói keresések során megadott keresőfeltételeket. A kétféle gyakorlat között annyi a különbség, hogy a keresőmotorok nem tudták, nem tudják jól kiválasztani a dokumentumot valóban jellemző, releváns tárgyszavakat (vagy pedig más relevanciakezelő megoldást alkalmaztak, mint a Google).

A probléma megoldására, a hiányzó szűrési, kiválasztási tevékenység elvégzésére szakemberek munkába állítása látszott megfelelőnek. Részben ezért indítottak a web kezdetén olyan szolgáltatásokat a szabadszavas keresőmotorok (a HotBot, az AltaVista és társaik) megjelenésével párhuzamosan, amelyek szemantikai szempontból kívánták meghaladni a keresőmotorok szolgáltatásait. A legismertebb próbálkozás a Yahoo Directory webkatalógusa volt, amely a weboldalakat egy saját osztályozási rendszer segítségével próbálta meg szemantikalag elrendezni és a felhasználók számára megtalálhatóvá tenni. Azonban a „drámát", az új világ fordulatát is ugyanennek a szolgáltatásnak a sorsában érhetjük tetten. A Yahoo egésze idővel ugyan a legsikeresebb webes szolgáltatások közé került, de a webkatalógusa egyre inkább háttérbe szorult, míg végül „bezárták", s a könyvtárosaikat, archivátoraikat szélnek eresztették (vagy más feladatokra irányították őket).[9]

A szakszerű és fegyelmezett rend reprezantánsa eltűnt, ám ezzel párhuzamosan megjelent valami más. A web2 jelenségkörbe tartozó szolgáltatások (mint a Flickr, Del.icio.us, Digg, YouTube stb.) ugyanis olyan metaadat-kezelési módszereket építettek ki maguknak, amelyek ugyanazt a munkát, amit addig a hagyományos és digitális archívumokban egyaránt szakemberek végeztek, az új szolgáltatásokban a felhasználók önkéntes munkájára bízták. Ezt a megoldást, pontosabban az ilyen rendszereket nevezték el *folkszonómiának*. A továbbiakban ezt jelenséget, illetve a folkszonómiák megjelenéséhez köthető paradigmaváltást vizsgáljuk meg néhány szempont alapján.

Elemzésünkben leegyszerűsített gondolatmenetet követünk a korábban felvázolt fogalmi modell összetevőire támaszkodva. A legfontosabb kérdésünk az lesz, hogy a digitális archívumok elterjedésével milyen módon lehet biztosítani a tárolt dokumentumok metaadatokkal történő ellátását és a dokumentumok visszakereshetőségét.

Legelőször a tudásszervezési rendszerek típusairól kell pár szót ejtenünk. A különböző archívumépítési gyakorlatokban a hálózati kultúra időszakát megelőző évszázadban háromfajta tudásszervezési rendszert vettek használatba:

- terminuslista
- taxonómia
- tezaurusz

Ezek mindegyike a rá jellemző matematika struktúrával írható le a legpontosabban. A tudásszervezési rendszerek alaphalmaza (D) azokat a szavakat, kifejezéseket (terminusokat) tartalmazza, amelyeket a dokumentumokhoz lehet rendelni. A tudásszervezési rendszerek különbségeit – első körben – az a tény határozza meg, hogy milyen – szintaktikai, szemantikai vagy más – relációkat (R_i) engedünk meg felvenni a rendszer elemei, illetve más adatok között.[10]

[9] AYahoo directory szolgáltatáshoz hasonló utat futott be a Google által felkarolt DMOZ Open Directory projekt is, amely szinte a kezdetektől fogva elég hosszú ideig elérhető volt a Google kezdőlapjáról, aztán egyszercsak lekerült onnan.

[10] Itt most nem térünk ki a relációk jellemzésére, tipizálására, de jelezzük, hogy a MEO-projekt dokumentumai közül több is foglalkozik ezzel a kérdéssel [MEO 2006], illetve a nyelvészek között részben ezzel foglalkozik [Cruse 1986] és [Lyons 1977]

Nem elég azonban csak a struktúrára figyelnünk, ha igazán meg akarjuk érteni a folkszonómiák jelenségét. Azt is fel kell vennünk a tudásszervezési rendszerek jellemzői közé, hogy van-e, s ha igen, milyen felügyelet, milyen kontroll van a metaadat-hozzárendelési munka folyamatában. Előbb persze meg kell mondanunk, miért is van szükség ennek figyelembe vételére. Nos, ha a tartalmi metaadatok dokumentumokhoz rendelésének az a fő funkciója, hogy egyértelműen jellemezni tudjuk velük a dokumentumok tartalmát, akkor az egyértelműséget (vagyis mindazoknak a többértelműségeknek az elkerülését, amelyeket a szabadszavas keresés gyöngeségei kapcsán felsoroltunk) biztosítanunk kell valahogy. A többértelműségek elkerülése pedig megfelelő szakértelmet, fegyelmezett munkarendet, szakmai kontrollt, kontrollált szótárakat, tudásszervezési rendszereket kíván. Ezt persze nem olyan könnyű formalizálni, hiszen olyan kérdésekre kell tudnunk válaszolni e szempont alapján, hogy:

Q_1 = támasztanak-e bármilyen szakmai feltételt, szaktudáselvárást a munkát végző személyekkel szemben?

Q_2 = van-e bármilyen munkaszervezési szabályrendszer, ellenőrzési mechanizmus a munka menetére vonatkozóan, azaz kik, milyen jogosultságokkal vehetnek részt a munka egyes részfolyamataiban?

Q_3 = kik rendelhetik a KOS-rendszer elemeit a dokumentumokhoz?

Q_4 = kik szerkeszthetik, módosíthatják, bővíthetik a tudásszervezési rendszer elemeit, relációit?

A tudásszervezési rendszerek építésének és alkalmazásának kontrolljára vonatkozó fenti Q_i kérdésekre különböző válaszokat adhatunk, és ezt a feltételegyüttest vagyis az S_i *társadalmi normák* összefüggő rendszerét érdemes felvenni a tudászervezési rendszerek jellemzői közé. A dolgokat kissé leegyszerűsítve a következő tevékenységekre vonatkozó *normákat* kell rögzítenünk:[11]

S_1 – kinek szabad új tárgyszót létrehozni a tudásszervezési rendszerben

S_2 – kinek szabad új relációt létrehozni a tudásszervezési rendszerben

S_3 – kinek szabad két tárgyszót relációba állítani a tudásszervezési rendszerben

S_4 – kinek szabad tárgyszót dokumentumhoz rendelni a katalógusban

S_5 – csak a tudásszervezési rendszer elemeit szabad-e a dokumentumokhoz rendelni

Az öt norma közül az első három a tudásszervezési rendszerek, az utolsó kettő a katalógusok építésével kapcsolatos. A számítógépek világában a fenti normák mind kezelhetők azáltal, hogy a digitálisan szabályozzuk, kinek van írási joga a tudásszervezési rendszer és/vagy a katalógus elemeire, illetve milyen adatokat lehet egymással összekapcsolni.
Az írási, szerkesztési jogosultságokat is figyelembe véve már felírhatjuk a tudásszervezési rendszerek teljesebb formuláját:

$KOS = < D, R_1, R_2, \ldots, R_n, S_1, S_2, S_3, S_4, S_5 >$, ahol

D a tudásszervezési rendszer tárgyszavaiból álló tartóhalmaz
R_i az elemeken (tárgyszavakon) értelmezhető reláció (i=1, …, n)
S_j a tudásszervezési rendszer társadalmi környezetét adó szabályrendszer (j=1, …, 5)

A fent bemutatott összetevőkkel már adott az az általános keret, amelyre támaszkodva elég pontosan megragadhatjuk a történelmileg létező, szélesebb körben elterjedt tudásszervezési rendszerek legfontosabb jellemezőit. Minden tudásszervezési rendszerben van egy közös reláció, a *lexikografikus rendezés*, amely a tárgyszóhalmaz elemeinek ábécé szerinti sorba állítását végzi.

A *terminuslistáknak* van a legegyszerűbb szerkezetük. Ilyenek a könyvek végén található indexek (név- és tárgymutatók), melyek a könyv legfontosabb kulcsszavait sorolják fel ábécé szerint, minden kulcsszóhoz hozzárendelve az oldalszámokat, ahol a kulcsszavak a szövegben előfordulnak, de ide tartoznak a különféle egységesített névlisták, sőt, a keresőmotorokban használt ‚invertált index' technológiája is. A terminuslista formulája a következő:

[11] A társadalmi normák formalizálásáról, típusairól lásd: [syi 2007]

KOS_{list} = < D, R_1, R_2, S_1, S_4, S_5 >, ahol

R_1 lexikografikus rendezési reláció

R_2 ekvivalenciareláció

A terminuslista esetében az S_2 és az S_3 szabály nem érvényesíthető, hiszen ebben a rendszerben új relációt nem lehet definálni. Az ilyen rendszernek a lexikografikus rendezés mellett van még egy másik relációja: a lista elemeit az R_2 *ekvivalenciareláció* kapcsolja össze. Egy földrajzi nevek egységesített besorolási rendszerét például az kapcsolja össze egyetlen egésszé, hogy minden tételéről azt állítjuk, hogy ekvivalensek egymással abban a tulajdonságukban, hogy valamennyien földrajzi entitások tulajdonnevei. Az ekvivalenciareláció fenntartásával azt kell „garantálnunk", hogy a terminuslista elemei – az alkalmazott szempont szerint – azonos minőségűek lesznek (tehát nem keverednek különböző típusú elemek, mondjuk személynevek a földrajzi nevekkel). Más relációt nem lehet a terminuslista elemei közé felvenni. Attól függően, hogy milyen típusú terminusokról van is szó, változhat az a gyakorlat, hogy fenntartják-e a a szavak, kifejezések bekerülését szabályozó S_1 normát. Az igazán komolyan vett két előírás a katalógusépítésre vonatkozó, vagyis a tárgyszavak és a dokumentumok összekapcsolását szabályozó S_4 és S_5 norma.

A *taxonómiák* (más néven *osztályozási* vagy *klasszifikációs rendszerek*) már két fontos szemantikai relációt tartalmaznak (a lexikografikus rendezés – „kötelező" – szintaktikai relációján túl). Ez a tudásszervezési rendszer úgy van felépítve, hogy az elemei hierarchikus módon egymás alá vannak rendelve – valamilyen *tartalmazási reláció* alapján. Ezt az alárendelési relációt lehet tiszta és pongyola értelmezés mentén is használni (a tiszta értelmezés esetben az alárendelési reláció a *generikus alárendeltje* relációval egyezik meg, a pongyola megközelítés keverten alkalmazza a generikus és a *partitív*, esetleg még más egyéb hierachikus relációt, például az *előzménye* viszonyt). A *hierarchia* leírásához azonban nem elégséges egyetlen relációt értelmeznünk a rendszeren, noha a közvélekedés gyakran megelégszik ezzel a megoldással. Arra is szükség van, hogy egy második relációval biztosítani lehessen azt, hogy az azonos felettes elem alá rendelt elemek különbözzenek egymástól, vagyis definiálni kell egy *különbözőségi relációt*.[12] A taxonómiát így a következőképpen írhatjuk le:

KOS_{tax} = < D, R_1, R_3, R_4, S_1, S_3, S_4, S_5 >, ahol

R_1 lexikografikus rendezési reláció

R_3 hierarchikus alárendeltje (tartalmazási) reláció

R_4 különbözőségi reláció

Ebben a szisztémában az S_i szabályok közül már négyet érvényesítenek, csak az S_2 norma hiányzik, hiszen nincs mód a két szemantikai kapcsolaton túl más relációt alkalmazni a rendszerben, amiből következően nincs is szükség az új relációk felvételét szabályozó normára.
A taxonómiák a könyvtári világ legelterjedtebb tudásszervezési rendszerei, az egész világon ezeket használják a könyvek tartalmának leírására (a Magyarországon használatos ETO-rendszer mintája és eredetije a Melville Louis Kossuth Dewey által kidolgozott Dewey Decimal Classification, DDC-rendszer). Népszerűsége az egyszerű kezelhetőségében rejlik. Ez az egyszerűség persze viszonylagos. A terminuslistákhoz képest ugyanis itt már szemantikai elvárásokat kell figyelembe vennünk, hiszen a hierarchikus alárendelési reláció alkalmazása (akármi legyen is az értelme egy konkrét taxonómia esetében) mindig szemantikai kényszerek betartásával kell, hogy együttjárjon. Ezért van az, hogy ezen rendszerek használata esetében már megkövetelik valamilyen szaktudás létezését és a munkafolyamat menetét is szabályozzák. Utóbbi mozzanat több részre osztható. A taxonómia elemeinek halmazát felfoghatjuk olyan *kontrollált szótárként* is, amelynek elemeit nem tetszőleges módon, hanem csak adott szabályokhoz igazodva, tehát csak kontrollált módon lehet bővíteni. Ez egyfelől korlátot jelent a tárgyszó-hozzárendelési munka során, mert előírásokhoz igazodó, tehát *fegyelmezett* munkavégzést követel meg az erre a feladatra előzetesen felkészített, *képzett archivátoroktól*, másrészt az ilyen rendszernek szüksége van egy olyan folyamatra, amely során a folyamatos változtatási igényeket ki lehet elégíteni, vagyis bővíteni, módosítani kell a rendszer valamely részét. Ez azt is jelenti, hogy szükség van *taxonómiaépítő szaktudásra* a rendszer fenntarthatósága végett. Ezek a feltételek, pontosabban az ezek teljesülésére vonatkozó kérdés azonban felvet két újabb, nagyon fontos tudásszociológiai, tudományfilozófiai kérdést:

[12] Lehetne még erősebb feltételt is előírni és a JEPD-elv teljesülését elvárni. Ez annyival több a közvetlenül függő elemek különbözőségének elvárásától, hogy azt is megköveteli, hogy az azonos szinten levő fogalmak „együttes terjedelme" megegyezzen a fölöttes elem terjedelmével. A JEPD-elv (jointly exhaustive and pairwise disjoint) magyar fordítása ‚együttesen kimerítő és kölcsönösen kizáró' lehetne. Bővebben lásd: [Bittner et al. 2004]

Q_5 = mennyire egységesen értelmezi az osztályozó közösség a taxonómia elemeit?

Q_6 = lehet egyetlen egységes rendszerbe rendezni valamely dokumentumgyűjteményt jellemző tudásterület fogalomkészletét?

A kérdésekre adott válasz szétfeszítené jelen tanulmányunk kereteit, úgyhogy a részletes kifejtéstől itt eltekintenünk, azonban egyetlen észrevétel felidézésével jeleznénk, hogy milyen irányban lehetne továbbszőni gondolatmenetünket e témában. Clay Shirky az ontológiák túlértékelt szerepéről írt cikkében tanulságos kritikát fogalmaz meg a DDC-vel szemben [Shirky 2005]. Miután megkérdezi, vajon miért van az, hogy a DDC-ben ugyanolyan fontosságot tulajdonítanak (azáltal, hogy azonos hierarchikus szintre helyezik őket) Ázsiának, Afrikának és a Balkán-félszigetnek, megadja a választ is: azért, mert nagyjából azonos számban adtak ki könyvet Amerikában a három földrajzi régióról, tehát a könyvtári polcokon elfoglalt helyigényük alapján tekinthetők ezek egyenrangú kategóriáknak. Akármennyire is jogos és elfogadható szempont ez a könyvtári világ számára, annyi azért kijelenthető, hogy a szempont elfogult. Ami után feltehető a kérdés, vajon lehet-e elfogultság nélkül tudásszervezési rendszert építeni, s a vélhető válasz az, hogy nem nagyon. Minden tudásszervezési rendszernek létjogosultsága lehet adott tudásterületen, az arra jellemző elfogultságokat figyelembe véve, de bajos olyan rendszert építeni és feltételezni, amely univerzális igénnyel léphetne fel, vagyis azzal a céllal, hogy minden tudásterületen, minden alkalmazási célra egyaránt felhasználható legyen. Mindez persze általánosítható, és nem csak a taxonómiákra, de a tezauruszokra is igaz. De lépjünk tovább, s nézzük meg, hogyan is tudjuk formalizálni az utóbbit.

Peter Mark Roget *tezaurusza* ugyanúgy a könyvtári világhoz tartozik, mint a taxonómia, a két rendszer nagyjából azonos időben jelent meg. Lényege az, hogy több és pontosabban rögzített relációt enged/követel meg a terminusok között. Kétféle – gyenge és erős – értelemben is lehet használni, mi itt az erős értelmezést mutatjuk be.
A taxonómiák hierarchikus alárendelési relációjának értelmezésekor általában megengedik azt, hogy az szemantikailag kevert legyen. Az ETO-ban például a hierachikus viszonyon legtöbbször olyan alárendelést értenek, amely a faja/neme viszonyt fejezi ki két elem között, ám olykor előfordul, hogy arra a fajta alárendelésre „használják“, amellyel az elemek közti rész-egész relációt ragadják meg – mondjuk az országoknak a kontinesek alá történő besorolásakor. A hierarchikus alárendeltje relációnak ezt a szemantikai többértelműségét igyekeznek kizárni a tezauruszok azáltal, hogy elkülönítenek szintaktikailag egyféle, de szemantikailag különböző hierachikus relációkat egymástól. Anélkül, hogy itt pontosan definiálnánk, a tezauruszok relációit csak felsorolásszerűen mutatjuk be. A tezaurusz rendszerét az alábbi összetevőkre bonthatjuk:

KOS_{tez} = < D, R_1, R_5, R_6, R_7, R_8, R_9, S_1, S_2, S_3, S_4, S_5 >, ahol

- R_1 lexikografikus rendezési reláció
- R_5 generikus alá- és fölérendeltje relációpár
- R_6 partitív alá- és fölérendeltje relációpár
- R_7 következménye-előzménye relációpár
- R_8 rokona (egyéb) reláció
- R_9 lásd/helyette szinonimareláció-pár

A tezaurusz esetében az S_i szabályok mindegyikét be kell tartani (az S_2-es szabályt nem mindig, sőt gyakrabban nem érvényesítik, vagyis nem engedik, hogy új relációt lehessen felvenni a rendszerbe, de a formális modellben azért kell felvennünk ezt a szabályt, mert előfordulhatnak olyan tezauruszok is, amelyekben a szabványokban rögzített relációkhoz képest további relációt is definiálnak). Vannak azonban a formulában jelzett relációk, relációpárok, amelyek más minőséget adnak a tezauruszoknak. Mivel a tezaurusz több, pontosan definiált relációt tartalmaz a taxonómiához képest, ezért összetettebb struktúrát képezhetünk le vele, és a gazdagabb szemantika, a nagyobb kifejezőerő miatt sokkal pontosabban, rugalmasabban és megbízhatóbban lehet vele a dokumentumok tartalmát leírni.
Ez az előny a kezdetektől predesztinálta arra a szerepre, hogy a tezaurusz váljék a könyvtári, archívumi világ legelterjedtebb tudásszervezési rendszerévé, ez azonban nem valósult meg. A tezaurusznak ugyanis az osztályozási rendszerekkel kellett viaskodnia, de a taxonómia-tezaurusz csatát már szinte az első pillanatban az előbbi nyerte meg. A csata kimenetét a könnyebb kezelhetőség döntötte el a javára. Mindez persze felveti azt, hogy egy alaposabb tárgyalás során figyelnünk kellene arra a szempontra is, hogy milyen kötelezettségei vannak a katalogizálást, metaadat-hozzárendelést végző embernek.
Ha a taxonómiák építéséhez kontrollált szótárra, kompetenciafeltételek fenntartására, fegyelemre, a munkafolyamatok ellenőrzésére van szükség, akkor ez még inkább így van a tezauruszok esetében, hiszen ott jóval bonyolultabb struktúrát kell fenntartani, több szempontra kell figyelni, nagyobb

szaktudásigényt kell elvárni a rendszert építőktől. Két – együttesen nehezen vagy egyáltalán nem teljesíthető – elvárás áll itt szemben egymással. Minél inkább szakterületi kérdésről van szó, annál megbízhatóbbnak lehet tartani a szaktezauruszok (szaktaxonómiák) használatát az adott tudásterület leírásában, azonban annál inkább szükség van szakképzett, fegyelmezett és ezért drága munkaerő alkalmazására. Ha pedig valami sokba kerül, akkor mindig felmerül a kérdés, hogy ki fogja megfizetni hosszútávon azt. A kontrollált rendszerek hanyatlásának magyarázatában ez a döntő mozzanat: egyre inkább az látszik, hogy a web kontextusában egyre kevésbé hajlandóak pénzt áldozni erre. Annál is inkább nehéz megfizettetni a kontroll árát, mert az utóbbi években megjelent új jelenség alternatív megoldás lehetőségét sejteti sokak számára.

A *folkszonómiák* jelensége és fogalma a web2-es szolgáltatásokkal együtt jelent meg. Maga a terminus Thomas Vander Wal egyik blogbejegyzésében bukkant fel először [Vander Wal 2004], de érdemes tudni azt a tényt, hogy az *etnoklasszifikáció* kifejezéssel Susan Leigh Star már 1996-ban nagyon hasonló értelmű fogalmat hozott létre [Star 1996]. Meg kell még itt említenünk, hogy ezeket a rendszereket gyakran *címkézési rendszerekként* (tagging system), olykor közösségi címkézési rendszerként is emlegetik, magát a metaadat-kezelési tevékenységet pedig címkézésnek hívják.
A folkszonómiák – a web2-es paradigmának megfelelően – a metaadat-hozzárendelési tevékenységet teljes mértékben saját felhasználóikra bízzák. Ezt természetesen csak akkor tehetik meg, ha nem várnak el semmilyen fegyelmet (és semmilyen speciális szaktudást) azoktól, akik a metaadatokat a dokumentumokhoz rendelik (azaz a felhasználói közösségük tagjaitól).[13] Ebből következően viszont a folkszonómiákat az eddigiekhez képest nagyon másként kell leírnunk:

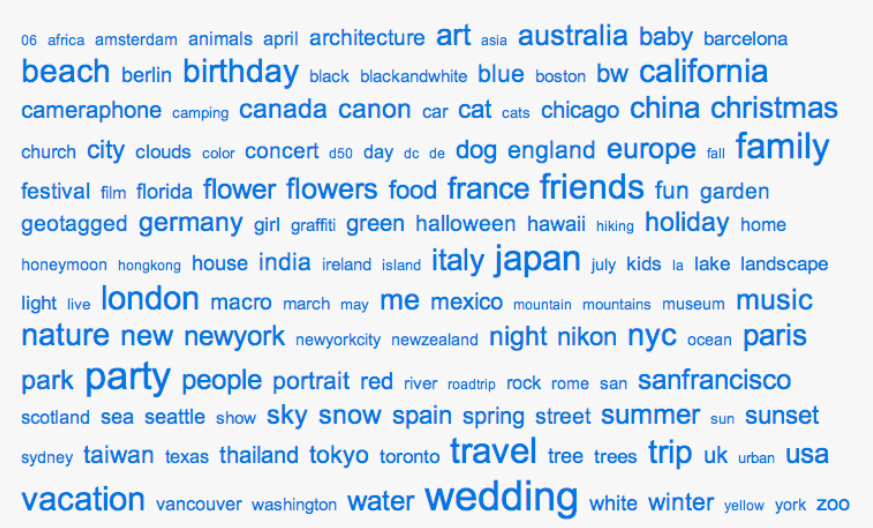
$KOS_{folk} = < D, T_D, R_1, R_{10}, R_{11}, R_{12}, S_4 >$, ahol

- T_D a D tartóhalmaz elemeire vonatkozó forgalmi adatok
- R_1 lexikografikus rendezési reláció,
- R_{10} ekvivalenciareláció
- R_{11} címkegyakorisági reláció
- R_{12} címkeegyüttjárási reláció

A folkszonómiák megjelenésével az a legfontosabb változás, hogy az S_i munkaszervezési szabályok közül csak egyet vesznek figyelembe (S_4-et). Azt sem mindig, sőt, talán még az is megkockáztatható, hogy nagyobb azoknak a folkszonómiáknak a száma, amelyek még ettől a szabálytól is eltekintenek (amikor figyelembe veszik az S_4-es szabályt, az akkor is „csak" annyit jelent, hogy a felhasználók kizárólag a sajátmaguk által feltöltött dokumentumokat címkézhetik, másokét nem).
Ebből a „kötetlenségből" sok minden következik. Ha nem követelünk meg semmilyen rendszerépítési szabályt, akkor egyrészt nem biztosíthatjuk a terminusok egyértelműségét a rendszeren belül (vagy másként mondva: nem lesz kontrollált szótárunk), másrészt szükségszerűen elveszítünk minden relációt a rendszerből, hiszen a szabad terminusfelvétel lehetősége mellett a címkézést végző személyektől nem követelhetjük meg azt, hogy az új címkéket hozzárendeljék a rendszerben már létező elemekhez. Ekkor viszont nem tudunk komolyabb struktúrát értelmezni a címkehalmazon, ami miatt nem számíthatunk a struktúra meglétéből fakadó – navigációs és következtetési lehetőségeket biztosító – előnyökre sem.
Fontos újdonságként minősíthetjük viszont a folkszonómiákkal megjelenő új – se nem szintaktikai, se nem szemantikai, sőt, nem is nyelvi, hanem forgalmi – relációkat (R_{10} és R_{11}).[14] Az R_{10} reláció egy olyan, a címkékhez rendelt gyakorisági viszony, amely azt mutatja, hogy a felhasználók milyen gyakran használják az adott címkét a dokumentumok leírására. Azért van ennek a relációnak kiemelt jelentősége, mert ezáltal újfajta relevanciakezelési lehetőséget lehet biztosítani a folkszonómiák számára. A címkegyakoriság ugyanis megmutatja az adott címkének a felhasználói közösségen belüli „népszerűségét", fontosságát, és ennek az értéknek a figyelembe vétele már elég jó alapot nyújt a relevanciakezeléséhez.[15] Másfajta segítséget képes nyújtani az R_{11} reláció, amely azt mutatja, hogy a többi felhasználó korábban milyen más címkéket rendelt a dokumentumokhoz az éppen használatban

[13] Nem is tehetnek mást. Ha a felhasználóktól bármit követelni akarnának, vélhetőleg mennének tovább olyan helyekre, ahol szabadon „mozoghatnak".
[14] Ez a két reláció tehát „kimutat" a tudászervezési rendszerből, hiszen ezek nem a tartóhalmaz elemei között vannak értelmezve, hanem a tárgyszavak és a rájuk vonatkozó forgalmi adatok között.
[15] A relevanciaképzésnek vannak más útjai is a közösségi szolgáltatások világában, bár sokuk már nem a folkszonómiákkal kapcsolatos felhasználói aktivitásra támaszkodik (például a szavazások, különféle rangsorok ilyenek, amelyek sokszor nagyon hasznosnak bizonyulnak).

levő címkével együtt. A címkéknek ezt a fajta együttjárását megmutatva a felhasználókat segíteni, orientálni lehet a megfelelő címkék megtalálásában.
A leggyakrabban használt címkék megjelenítését *címkefelhőnek* nevezik, amelyet úgy mutatnak be, hogy a címkék betűméretével vagy egy számértékkel jelzik a címkék népszerűségét, azok gyakorisági értékeit (lásd 2. ábra).



2. ábra: a Flickr címkefelhője, 2007.05.10.

Ha figyelmesen megnézzük a fenti ábrát, a folkszonómiák több komoly problémáját észrevehetjük rajta. A kontroll hiánya többféleképpen tetten érhető. Több olyan címkepár van, amely ugyanazon fogalom egyes- és többesszámú alakjára vonatkozik (‚cat' és ‚cats', ‚flower' és ‚flowers', ‚tree' és ‚trees'), amelyeket a kontrollált szótárakkal ki lehet szűrni. Mivel ezek a rendszerek a címkéket automatikusan detektálják (és szóközök közti karaktersorozatot tekintenek egy címkének), a több szavas kifejezéseket „feldarabolják", amit az emberi feldolgozás nyilván nem tenne meg. Példa lehet erre a ‚New York' vagy a ‚black & white' tárgyszavak (amelyek ebben a formájukban nem is szerepelnek a rendszerben, hiszen több szóból álló „kifejezések"), amelynek tagjai (a ‚new' és a ‚york', illetve ‚black' és ‚white') szerepelnek külön is, de egybeírva is (‚newyork' és ‚newyorkcity', illetve ‚blackandwhite' és ‚bw'). Még a ‚New York' példánál maradva az is látszik, hogy a folkszonómiák nem kezelik a szinonimitást sem, hiszen a ‚newyorkcity' és ‚nyc' címkék nyilván ugyanarra a fogalomra mutatnak, mégis külön szerepelnek a rendszerben. A címkefelhőből magából még nem látszik, de rövid idejű használat során könnyen felderíthető, hogy a folkszonómiák nem kezelik a szemantikai többértelműség többi fajtáját sem (homonima, poliszéma stb.).[16]
A felhasználói címkézést más szempontból is kritizálni szokták. Konkrét példaként hivatkozhatunk a fenti címkefelhő ‚me' címkéjére, amely a közösség számára nyilván kezelhetetlen kategória (ehhez hasonlóak még a ‚toread', vagy ‚todo' címkék). A címkék egy jelentős része személyes használatra való, de a közösség egésze számára értéktelen, használhatatlan. Az is gyakori jelenség a folkszonómiák gyakorlatában, hogy egyes felhasználók a többiek számára meglepő, gyakran érthetetlen címkéket aggatnak bizonyos dokumentumokra (például egy macskát, kutyát mutató videót a ‚bb' címkével ír le valaki), vagy nem kevés esetben a felhasználók hibás alakban adják meg a címkéiket (‚cat' helyett ‚cad' címkét rendelik a „macskás" dokumentumhoz).
Akármennyi hibát (mégpedig rendszerhibát) találunk is a folkszonómiák világában, mégis működőképesnek tűnik az egész. A nagylétszámú közösség tagjainak apró munkája „szervesül", a sokaság eltűnteti az egyének egyedi „hibáit". Erre utal az a megfigyelés, amely szerint a címkék relatív gyakorisága igen hamar beáll egy állandó értékre, vagyis a dokumentumokhoz rendelt címkék megoszlása stabilizálódik [Golder & Huberman 2006]. Ez annyit tesz, hogy a közösség egésze végül is konszenzusosnak mondható címkekészletet képes a dokumentumokhoz rendelni.
A szubjektív címkék azért nem okoznak igazán problémát, mert azok szerint úgysem akarnak keresni a felhasználók, így nem is zavarja őket az ilyen címkék jelenléte. Hasonlóképpen, a hibás alakok vagy az érthetetlen, egyéni címkék is „lesüllyednek címketenger mélyére", és nem igazán látszanak (tehát nem is zavarnak) a sokak által használt címkékhez képest.

[16] A YouTube-on például a ‚cat' címkére keresve sok „macskás" videót kapunk, de előfordulnak ‚Cat' nevű személyekhez kapcsolt anyagok is.

A szemantikus vakság problémája sem jelent akkora gondot a folkszonómiák esetében, mivel ha a keresési oldalon nem kapunk egyértelmű minősítést, megkülönböztetést a keresőfeltételek megfogalmazásakor (márpedig nem kapunk, hiszen a felhasználók nem adják meg az egyértelműsítéshez szükséges többletinformációt, amikor például beírják a ‚cica' keresőfeltételt), akkor a keresések feldolgozása során sem tudjuk igazán feloldani a nyelvi többértelműségeket.
A folkszonómiák terjedésének, dominássá válásának legfőbb okát abban látjuk, hogy a metaadat-hozzárendelés nehéz, fáradságos emberi munkáját sokak számára lehetővé téve, nagyszámú „szabad" és – ami a legfontosabb – „ingyen munkát" várhatunk a felhasználók önkéntes seregétől. Ahhoz, hogy ez működjön, nem lehet semmilyen megkötést előírni a címkézési munkát végzők számára, tehát fel kell adni az előzetes szakképzésre, szakértelemre vonatkozó elvárást, a konktrollált szótárak fenntartásának igényét és a munka ellenőrzésének lehetőségét.
Itt állunk tehát a kontroll elve alapján működő, a minőség ígéretét adó rendszerek folyamatos (és szerintünk megállíthatatlan) süllyedésénél az egyik oldalon, illetve az emergens fokszonomikus rendszerek felemelkedésénél a másik oldalon, és legalább hipotéziseket kéne tudnunk megfogalmazni arra, hogy mit várhatunk a jövőtől. A folkszonomikus rendszerek terjedésével ugyanis egyre nagyobb szemantikai rés keletkezik, amelyet valahogy át kellene hidalni szemantikai tudás alkalmazásával. A kérdés az, hogy milyen módon lehet a hiányzó tudást a rendszerbe betáplálni. Nem érdemes reménykedni abban, hogy az archivátori, könyvtárosi kompetencia és tudás hiányát a világ egyszer csak felismeri, és „visszahívja, rehabilitálja" őket. Más úton lehetne hasznosítani ezt az évszázados örökséget.
A weben keresztül elérhető dokumentumok egyre növekvő számával már a folkszonómiák sem bírják a versenyt. Ami lehetőség egyáltalán megmarad a szemantikai rés nagyságának csökkentésére, az a számítógép alkalmazása erre a célra. Meítélésünk szerint ez reális (bár a távoli jövőre vonatkozó) remény, hogy a számítógépeket megtanítva a szövegértésre, velük végeztetjük el a szövegek automatikus feltárását. Vagyis arra kell várnunk, hogy a szemantikus web kezdeményezés betöltse küldetését. Ehhez azonban biztosan nem a mérnökökön, pontosabban, nem csak a mérnökökön keresztül vezet az út. Nem egy giga tudástárat, nem egyetlen hatalmas ontológiát, hanem rengeteg, kontextusérzékeny tudástárat, szakontológiát kell felépítenünk, amelyek egyik kiemelt használati célja az lehet, hogy velük tanítani lehet a gépeket az automatikus szematikai feltárás munkájának végzésére. Ez pedig nem fog menni szakemberek, szaktudás és fegyelem, könyvtárosok, archivátorok nélkül.

Hivatkozások

[Berners-Lee et al. 2000]
Tim Berners-Lee, James Hendler, and Ora Lassila, The Semantic Web, in: *Scientific American*, May 2001, at: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

[Shirky 2005]
Clay Shirky, *Ontology is Overrated: Categories, Links, and Tags*, 2005
at: shirky.com/writings/ontology_overrated.html

[Golder & Huberman 2006]
Scott A. Golder, Bernardo A. Huberman, Usage Patterns of Collaborative Tagging Systems, in: *Journal of Information Science*, 2006, 32(2). 198-208.

[syi 2007]
syi, *Egyben az egész. Egytől egyig*, Typotex, 2007.

[O'Reilly 2005]
Tim O'Reilly, *What Is Web 2.0. Design Patterns and Business Models for the Next Generation of Software*, http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[MEO 2006] – http://ontologia.hu/meo

[Vander Wal 2004]
Thomas Vander Wal: *Off the Top*, 2004, at: http://www.vanderwal.net/random/category.php?cat=153

[Star 1996]
Susan Leigh Star, *Slouching toward Infrastructure*, 1996, at: http://is.gseis.ucla.edu/research/dl/star.html

[Cruse 1986]
D. A. Cruse, *Lexical Semantics*, Cambridge University Press, 1986

[Lyons 1977]
J. Lyons, *Semantics I-II*, New York: Cambridge University Press, 1977

[Bittner et al. 2004]
Thomas Bittner, Maureen Donnelly, Barry Smith : Individuals, universals, collections: On the foundational relations of ontology. In Achille C. Varzi, Laure Vieu (szerk.): Form*al Ontology in Information Systems. Proceedings of the Third International Conference (FOIS 2004)*, 2004, IOS Press, 37–48. p