Ungváry Rudolf
Országos Széchényi Könyvtár

A tartalom szerinti információkeresés az interneten

II. Internetkatalógusok

Az internetes keresőszolgáltatások mind rugalmasságban, felhasználóbarát felületek dolgában, mind az információs kínálatban messze felülmúlják a távolsági online szolgáltatások adta lehetőségeket. Mindez kihívás az információkeresés és osztályozás számára, amely az internet megjelenésével történetének legjelentősebb fejlődése előtt áll. A keresőszolgáltatásokat kezdettől fogva ugyanaz a kettősség jellemzi, mint minden hagyományos tartalom szerinti kereső és rendező rendszert: kialakultak a természetes nyelven működő, olykor már szabványosított szótárakat (tezauruszokat) is alkalmazó indexelőszolgáltatások, és a hierarchikus osztályozási rendszereket alkalmazó internetkatalógusok. Frissen kialakult szóhasználatukat megkíséreljük összehangolni a dokumentációs–könyvtári terminológiával. E második részben az internetkatalógusokkal foglalkozunk, végül röviden kitérünk az elsődleges és másodlagos elektronikus dokumentumok formátumaira is.

4.4 Internetkatalógusok

4.4.1 Meghatározás

Az internetkatalógusok (browsing services, browsing Dienste) hierarchikus (ritkábban enumeratív) osztályozási rendszert alkalmazó keresőszolgáltatások, melyek adatbázisa a túlnyomórészt intellektuálisan osztályozott HTML-dokumentumok rekordjait (másodlagos adatokból álló leírásait) tartalmazza, valamint egyéb adatbázisok információtételeit. Bennük az osztályok alapján – elsősorban a katalógusban „lapozva" – végezhető a böngészés.

Az ismertebb globális rendszerek közé tartozik például az Excite, Magellán, Northern Light, Yahoo!. A keresőszolgáltatásoknak ez a fajtája jelent meg először, valójában már a web előtt, a Gopherrel egy időben. Magyarországon 1995-től működik a HUDIR (Hungary.Network), 1999-től a Kincsereső/Kapu (Elender), 2000-től pedig az AltaVizslának (Matáv) is van az indexelőszolgáltatás mellett saját katalógusa.

Nevezik ezeket böngészőszolgáltatásnak, tárgyszótárnak, tématárnak (subject directories, Themenverzeichnisse).

4.4.2 Forráskiválasztás

A manuálisan előállított internetkatalógusokra jellemző, hogy kisebb-nagyobb mértékben intellektuálisan sorolják be (osztályozzák) a HTML-dokumentumokat az alkalmazott osztályozási rendszerbe. Automatikus osztályozással működő rendszerekből alig van néhány (ilyen a Gerhard és a Scorpion).

A feldolgozandó dokumentumok kiválasztását elvileg ugyancsak intellektuálisan végzik, de nagyon különböző színvonalon. A szolgáltatások egy részében semmiféle aktív kiválasztás nem zajlik, kizárólag olyan katalogizált tételeket tartalmaznak, amelyeket önkéntesen adnak át a honlapok tulajdonosai, szerzői, akik többnyire az osztályozásról is gondoskodnak, vagy legalábbis szabad tárgyszavakkal, tartalmi leírással látják el beküldött tételeiket.

A szolgáltatások többségében ugyan válogatnak, a kiválasztás kritériumai azonban alig ismerhetők meg. A különböző felmérések tanúsága szerint úgy fest, mintha a dokumentumok feltárását általában nem előzné meg határozottan körvonalazott gyarapítási tevékenység, csak afféle „spontán érkeztetés" zajlik.

„Maguk a szolgáltatások személyes megkérdezés esetén is csak nagyon kevés, illetve pontatlan információt közölnek kiválasztási kritériumaikról, a honlapjaikon pedig általában semmiféle tájékoztatás nem található róluk. Feltehető, hogy a kiválasztást sokszor nem valami tudatosan végzik, még ha olykor léteznek többé-kevésbé pontosan megfogalmazott követelmények. Többségükben szerkesztőket alkalmaznak, de nem ismerhető fel, miféle szelekciót végeznek: minden jel szerint nem annyira a kiválasztásra helyezik a hangsúlyt, mint inkább a tartalmi feltárásra. Egy tanulmányban a Yahoo!-ról ez szerepel:

Először összegyűjtik az új weboldalak URL-jeit. A legtöbb közülük drótpostán érkezik azoktól, akik a hálón szereplő oldalaikat szeretnék fölvetetni, a többit pedig a Yahoo! leszedője szállítja – egyszerű robot, mely új weboldalakat keresve hiperlinkről hiperlinkre ugrál. Ezt követően a húsz osztályozó valamelyike átnézi a weboldalt, és elvégzi a besorolást.

Különösen a nyelvi vagy tematikus alapon szelektáló szogáltatások esetén nincs információ a kiválasztáskor figyelembe veendő tartalmi kritériumokról. Legfeljebb azt említik, hogy félig üres weboldalak nem jöhetnek szóba, az UK Web Library (a brit „nemzeti" katalógus") pedig bizonyos tartalmú (pl. trágár) dokumentumokat kizár a gyűjtésből. Az általános gyűjtőkörű szolgáltatásokban az előbbiekhez képest inkább alkalmaznak tartalmi és formális kritériumokat.

A szerkesztőket alkalmazó szolgáltatásokban a döntéseket minden jel szerint intuitíve, a szakmai tapasztalatok alapján hozzák. (Magellan: Minden szerkesztőnk szakember a maga területén, ezért a végső döntés mindig az ő kezében van.) *Részletezett, konkrét kiválasztási kritériumokat a 12 általános és globális szolgáltatás közül csak az Argus Clearinghouse, a NetFirst és a Webcrawler select közölt.*

Részletesebben tájékoztattak a szolgáltatások a feldolgozott weboldalak minősítési (rating) kritériumairól (átfogó és egyedi tartalmi, megjelenési és technikai/szoftver minősítés).

Alig van olyan szolgáltatás, amelyben megkülönböztetnek feltétlenül betartandó és másodlagos kritériumokat, nem is súlyozzák ezeket. Az Argus Clearinghouse bizonyos metaadatok (szerzőség, dátum) létét elengedhetetlennek tekinti, a Lycos A2Z számára a más weboldalról származó hipercsatolók gyakorisága a legfontosabb kiválasztási feltétel.

Beszélni kell az itt felsorolt kritériumok operacionalizálásáról. Erről akkor van szó, ha a feltételeket mérhető adatokkal kapcsolják össze. Melyek konkrétan a kizárandó és a fölveendő tartalmak? Mennél nem régebbi weboldalak vehetők föl? Milyen metaadat megléte elengedhetetlen stb. Az objektív felhasználhatóság érdekében az arra alkalmas kritériumokat operacionalizált formában kell megfogalmazni. A weboldal látogatási gyakoriságának, idézettségének (hipercsatoltságának) megkövetelt határértékeit például számszerűen is meg kell adni. Vizsgálatunkban a 19 megkérdezett szolgáltatás közül egyetlenegy sem említett operacionalizált feltételeket" [10].

A kritériumok a vizsgálatok alapján az alábbiakban foglalhatók össze (az aláhúzottak a feltétlenül betartandók, a többiek másodlagosak):

1. Stabilitási kritériumok:
   - 1.1 a forrás könnyen és biztosan elérhető
   - 1.2 a forrás előreláthatólag nem rövid életű
   - 1.3 a forrás aktualizálására, karbantartására számítani lehet
2. Tartalmi kritériumok:
   - 2.1 a forrás tartalma hihető, létrehozója a tartalom vonatkozásában hiteles, megbízható testület vagy személy
   - 2.2 a forrás időszerű
   - 2.3 a forrás érdekes, közérdeklődésre tart igényt
   - 2.4 a forrás informatív, érdekes
   - 2.5 a forrás jól szerkesztett, részletes, egyedi, tipikus, speciális
   - 2.6 a forrás nem tartalmaz olyasmit, ami a mindenkori kizáró tényezők jegyzékében szerepel
3. Formai kritériumok:
   - 3.1 a forrás nem régebbi, mint ...
   - 3.2 a forrásnak megvannak a felsorolt metaadatai (cím, szerzőség/közreadó, tárgyszavak), HTML-szerkezete szabványos
   - 3.3 a forrásban sok más forrásra vonatkozó csatoló van, különösen a teljes HTML-dokumentumokra, szolgáltatásokra utal
   - 3.4 a forrásra gyakran utalnak más forrásokból
   - 3.5 a forrást gyakran használják, sok a látogatója
   - 3.6 a forrás nem túl kicsi (hacsak nem nagyon időszerű, közérdekű)
   - 3.7 a forrás szép, látványos, különleges formatervezésű
   - 3.8 a forrás ingyenes

4.4.3 Avulás és frissítés

Az internetkatalógusok állományai ugyanúgy avulnak, akár az internet többi állománya. Frissítésükre azonban még az indexelőszolgáltatásokban alkalmazott gyakoriságoknál is ritkábban kerül sor, mivel a katalógusok HTML-dokumentumait intellektuálisan dolgozzák föl, s nem mindig áll rendelkezésre olyan keresőgép, amely a frissítést végrehajthatná. Ezért az internetkatalógusokban sokkal több a zsákutcás HTML-rekord (dead link), amelyből kiindulva az eredeti HTML-dokumentum már nem hívható elő.

4.4.4 Osztályozási rendszerek

4.4.4.1 Hagyományos osztályozási rendszereket alkalmazó internetkatalógusok

McKiernan, az iowai egyetem könyvtárosának mutatója, a *Beyond Bookmarks* [1], amely a hagyományos osztályozási rendszereket, tárgyszójegyzékeket és tezauruszokat használó keresőszolgáltatásokról tájékoztat,* 1999 végén 55 olyan internetkatalógust sorol föl, amelyben hagyományos osztályozási rendszereket használnak. Ezen belül 22 a Dewey Tizedes Osztályozását, 11 az ETO-t és 6 a Kongresszusi Könyvtárét.

A dokumentációs–könyvtári, vagy egyéb bevált hagyományos osztályozási rendszer alkalmazása elsősorban azokra a szolgáltatásokra jellemző, amelyek fölhasználói köre tudományos és egyéb szakemberekből áll, és ezért elsősorban tudomá-

* Egy másik ilyen mutatót a DESIRE projekt tartalmaz [8]. Egyszerűbb összeállítás a düsseldorfi egyetem könyvtárosának, *Barbara Lutes*nak a *Thesaurus compendiuma*, amelyben nem az interneten használt, hanem közvetlenül vagy közvetve elérhető tezaurusz, osztályozási rendszer, illetve csak annak nevezett információkereső nyelvi szótár csatolóit gyűjtötte össze.

nyos jelentőségű forrásokat dolgoznak fel. A feldolgozás kiválasztási kritériumainak itt lényegesen nagyobb a jelentősége. A hagyományos osztályozási rendszereket többnyire kisebb internetkatalógusok használják, egy részüket a könyvtárak hozták létre (pl. BUBL, NISS, WWW Virtual Library, NetFirst).

A hagyományos, bevált és tudományos igénnyel készült osztályozási rendszerek alkalmazóin belül külön csoportot alkotnak azok a szakterületekre specializálódott gyűjtőkör-katalógusok, amelyekben minőségbiztosítási szempontokat alkalmaznak a kiválasztásban és feldolgozásban, részletes tartalmi és formai leírást készítenek, többek között annotációt, összefoglalásokat, és a munkákat a szakterület szakértőivel végeztetik el. Ezeket *szakterületi információs kapuszolgálat*oknak (subject based information gateway) nevezik. Pl. az informatikai weboldalakat feldolgozó Ariadne, amelyben az ACM számítástechnikai osztályozási rendszerét (Computer Classification System), vagy az Engineering Electronic Library System (EELS), amelyben speciális osztályozási rendszert és az EI tezauruszt használják.

Ebben a körben jelennek meg az automatikus osztályozást alkalmazó internetkatalógusok is: Scorpion, Gerhard (részletesen beszámol róluk [7]).

4.4.4.2 Önállóan kialakított osztályozási rendszert alkalmazó internetkatalógusok

Ezek alkotják az internetkatalógusok túlnyomó többségét.

A legfelső szinten néhány jól áttekinthető, és főleg közismert szakterület (főosztály) jelenik meg. Az osztályozási rendszerek többnyire ismeretterületeket tartalmaznak, de vannak földrajzi, időrendi, dokumentumtípusok stb. szerinti rendszerek is.

A nagyobb, nemzetközi internetkatalógusokban szinte mindenütt saját fejlesztésű egyetemes osztályozási rendszereket használnak, melyeket túlnyomórészt a hagyományos osztályozási rendszerektől teljesen függetlenül, feltehetően azok ismerete nélkül, elsősorban kereskedelmi szempontokat figyelembe véve alakítottak ki. A főosztályok kiválasztása és rendezettsége messzemenően a köznapi nyelvhasználat, gondolkodás és tájékozódás igényeit tükrözi. Ez egyben friss látásmód is az osztályozási rendszerek alapvetően konzervatív világában, és előbb-utóbb számolni lehet megtermékenyítő hatásával a könyvtári-dokumentációs osztályozásra. Ugyanakkor számtalan következetlenség, dilettantizmus és rövidlátó prakticizmus forrása. Ezekben az osztályozási rendszerekben olykor rendkívül rugalmasan alkalmazott megoldásra bukkanunk, jelentős részük a web körülményei között akkor is beválik, ha logikailag ellentmondásos, de gyakoriak a rendszer koherenciáját gyengítő megoldások is, amelyek a későbbi fejlődés során bonyodalmakat okozhatnak.

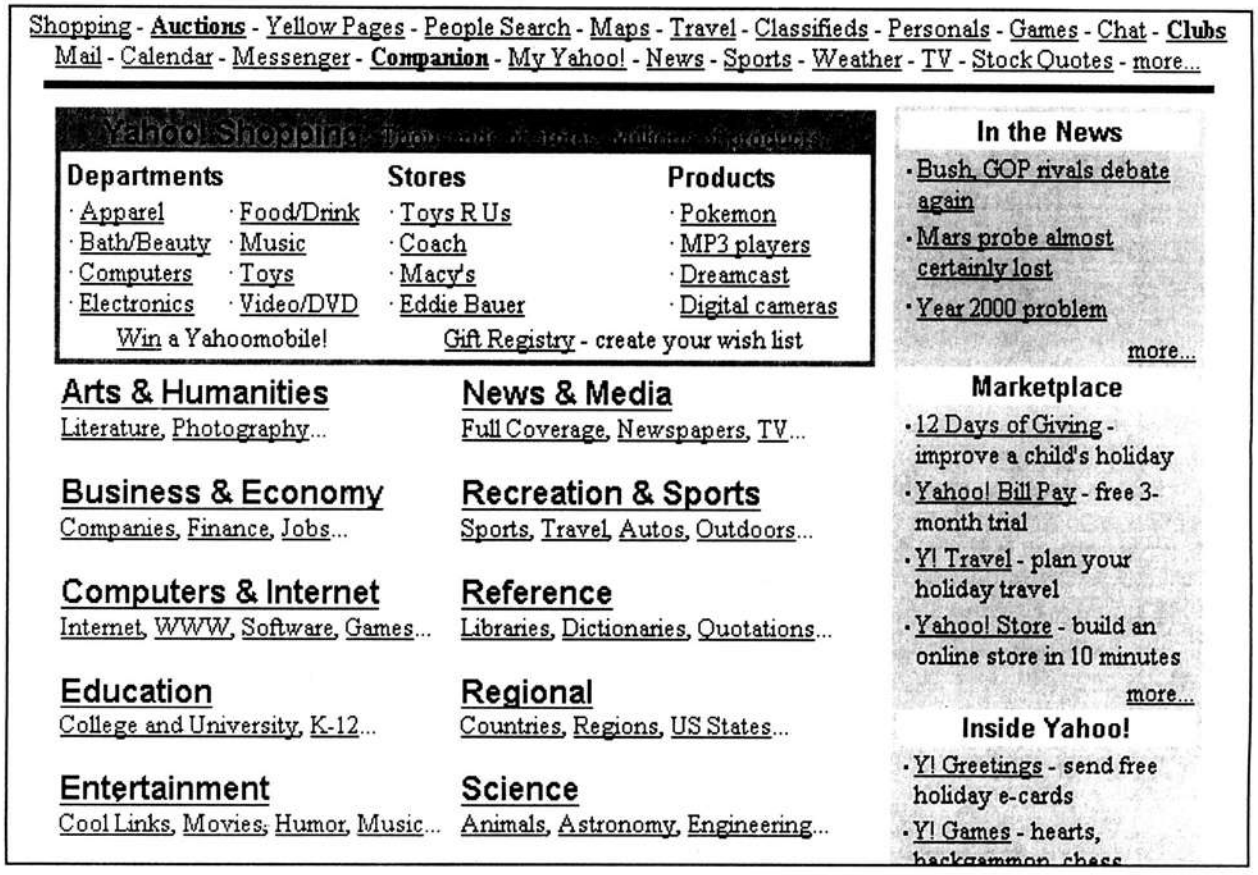
Az *1. ábrán* az egyik legismertebb internetkatalógus, a Yahoo! kezdőlapján megjelenő osztályozási rendszer legfelső hierarchiaszintje látható.

A nagy keresőszolgáltatások ma mintegy internetes húzóágazatként működnek, jelentőségüket nem lehet eléggé felbecsülni. Egyetemes igényű osztályozási rendszereiknek futtában végzett készítési és fejlesztési körülményeire fényt vet az alábbi interjúrészlet, amelyben a Yahoo! osztályozási rendszerének szerzője a következőket nyilatkozza:

„Négy hónappal ezelőtt Srinavasan közölte velem, hogy további kategóriákat vett föl, és szinte minden nap változtat valamit az ontológián" [14].

Az internetkatalógusok osztályozási rendszereinek osztályait – függetlenül azok szintjétől – a szolgáltatók általában „kategóriáknak" nevezik. Ez, és sok más elnevezésbeli eltérés a hagyományostól feltehetően éppen abból ered, hogy a készítőkben nem is tudatosult: olyan rendezőrendszert terveztek és használnak, amelynek osztályaiba besorolják az információtételeket, azaz a rendszer segítségével osztályoznak. Innen nézve nem a rendszer logikai/filozófiai (kategoriális), hanem besoroló, „tartalmazó" szerepéről van szó, azaz dolgok (HTML-rekordok) osztályairól (nem pedig HTML-rekordok „kategóriáiról"). Az osztályozási rendszer sem „ontológia", noha ugyanúgy létezik, akár a sertéscsülök, mivel az ontológia (a létről szóló tan) a filozófia egyik ága, tehát tudomány, az osztályozási rendszer viszont nem tudomány, hanem konkrétan létező termék. A hierarchikus osztályozási rendszerek korántsem olyan „nyitottak", mint a tárgyszójegyzékek vagy tezauruszok, s ezért teljesen alkalmatlanok arra, hogy konzisztenciájuk összeomlása nélkül naponta változtatgassanak rajtuk.

A tervezők osztályozási hagyományoktól való érintetlensége abban is megmutatkozik, hogy az egyes szinteken az ilyen típusú rendszerek többségében az osztályokat nem szisztematikusan, hanem betűrendben jelenítik meg. Indokaik kétségtelenül nyomósak: a lehető legkevesebb szellemi erőfeszítést szeretnék okozni a végfelhasználónak. A legfelső szinten még nem annyira feltűnő, hogy a hierarchikus rendszer adott szintjén a betűrend miatt össze nem tartozó osztályok kerülnek egymás mellé, mert ezen a szinten minden keresőszolgáltatásban a lehető leggyorsabb áttekintésre törekszenek: egy pillantással lehessen fölmérni, hogy a rendszer lényegében mit és hol tartalmaz. Az alsóbb szinteken azonban szokatlan találkozások adódnak. A Science (Tudomány) második szintjének több mind 60 osztálya például így kezdődik: Acoustics (Akusztika), Agriculture



1. ábra **A Yahoo! internetkatalógus belépőlapjának részlete, melyen az osztályozási rendszer legfelső szintje látható**

Az elválasztó vonal fölött az osztályozási rendszer hierarchiájától elkülönített osztályok kifejezései láthatók, melyek egy-egy adatbázis (pl. Shopping [Bevásárlás], Classifieds [Apróhirdetések]) vagy szolgáltatások (pl. My Yahoo! [a Yahoo! átszabása személyes igényeknek megfelelően]) belépőpontjai.

(Mezőgazdaság), Alternative (Alternatív technikák), Amateur science (Amatőrök által művelt szakterületek), Antropologhy and Archeology (Embertan és régészet), Artificial Life (Mesterséges élet) stb.

A hierarchikus rendszer nem különösen „mély": alig 3-4 szintet tartalmaz. Ezért jelenik meg a második és a harmadik szinten olykor nagyon sok osztály. A szerkesztők valószínűleg nem mernek a már széles körben megismert főszerkezeten változtatni; ilyen változtatás nélkül azonban nem oldható már meg, hogy az egyes szinteken az osztályok számát csökkentsék. Az egész emlékeztet a termeszek hangyabolyépítményeire: a fejlődés szerves és nagyon gyakorlatias, mindig kizárólag a lehetőségekhez igazodik, sohasem elvekhez. Kétségtelen, hogy az elvek alkalmazásának vannak praktikus határai. De az is igaz, hogy a prakticitás túlfeszítéséből is adódnak határok. Van, amikor már nincs megtévesztőbb, mint a realitás.

Az eddig megjelent átfogó internetkatalógusok egyetemes célú osztályozási rendszereit nem jellemzi a felosztási szempontok következetessége. Érezhető, hogy kereskedelmi szempontok érvényesülnek az osztályok fölvételében: az a felfogás, hogy „mi van azon a szakterületen eladható információ". Ez határozza meg, milyen osztályokat vesznek föl a rendszerbe. Csak feltételezzük, hogy a keresőszolgáltatások gépei által feldolgozott információtételek mennyiségének növekedésével a rendszerek finomszerkezete tartalmilag fokozatosan koherensebbé válik. Ugyanakkor az alkotók szakmai érintetlenségének előnyei is vannak: friss szemmel vágtak neki a világ rendszerező célú felosztásának, s ez hosszabb távon nem maradhat következmények nélkül a hagyományos könyvtári és dokumentációs osztályozásra sem.

Különösen hasznos megoldások születtek az ilyen osztályozási rendszerek hierarchialáncai között. Ennek alapja, hogy a hipertext *a* kereszthivatkozások eszményi rendszere, és ezt hasznosítják a hierarchikus szerkezeten belül is. Itt is létrehoznak keresztirányú összefüggéseket. Ez abban nyilvánul meg, hogy egy-egy osztály egyszerre több magasabb szintű osztály alárendeltje is lehet, az osztályozási rendszerek tehát – szemben a hagyományos egyetemes könyvtári rendszerekkel – polihierarchikusak. Ez olykor rendkívül bonyolult,

néha már lehetetlennek tűnő struktúrákat eredményez, de a felhasználót nagyon jól szolgálja, mert az ismétlődések következtében a hierarchikus rendszer redundáns.

A *2. ábrán* azt láthatjuk, hogy például a Motorcycles (Motorkerékpárok) hány különféle hierarchialáncon belül jelenik meg. Mindig van „gazdaosztály" („szülőosztály"), amelyhez a polihierarchikusan alárendelt alosztály kapcsolódik (a többi előfordulást a megjelenítésben a @ jellel jelölik).

ul a Motorkerékpárok osztályai között vannak olyanok, amelyek a Recreation főosztály fokozatos alosztásaiból keletkeztek. A „Recreation–Automotive–Motorcycles" és a „Recreation–Hobbies–Models–Motorcycles" láncban a Motorkerékpárok osztálya nem ugyanaz az osztály-előfordulás a rendszeren belül, mint mondjuk a Bussines and Economy–Companies–Automotive–Motorcycles láncban szereplő Motorkerékpároké. Ezért az előbbi két osztálylánc Motorkerékpárok osztályát a

Yahoo! Category Matches (1 - 20 of 24) SHOP-Ame

Recreation > Automotive > **Motorcycles**

Business and Economy > Companies > Automotive > Shopping and Services > **Motorcycles**

Business and Economy > Companies > Automotive > Business to Business > **Motorcycles**

Recreation > Automotive > **Motorcycles** > Vintage **Motorcycles**

Net Events > Recreation > Automotive > **Motorcycles**

Recreation > Automotive > **Motorcycles** > Feet Forwards **Motorcycles**

Business and Economy > Companies > Financial Services > Insurance > Automotive > **Motorcycles**

Recreation > Hobbies > Models > **Motorcycles**

Business and Economy > Companies > Automotive > Shopping and Services > **Motorcycles** > Makers > Honda **Motorcycles**

2. ábra **A Motorkerékpárok (Motorcycles) polihierarchikus előfordulása a Yahoo! osztályozási rendszerében**

A *3. ábrán* a Motorkerékpárok osztály alatti utolsó előtti hierarchiaszint látható. Megjelenítettük az első néhány találatot is azok közül az információtételek közül, amelyeket az átfogó Motorkerékpárok osztályba soroltak, és nem az ennél speciálisabb alosztályok valamelyikébe.

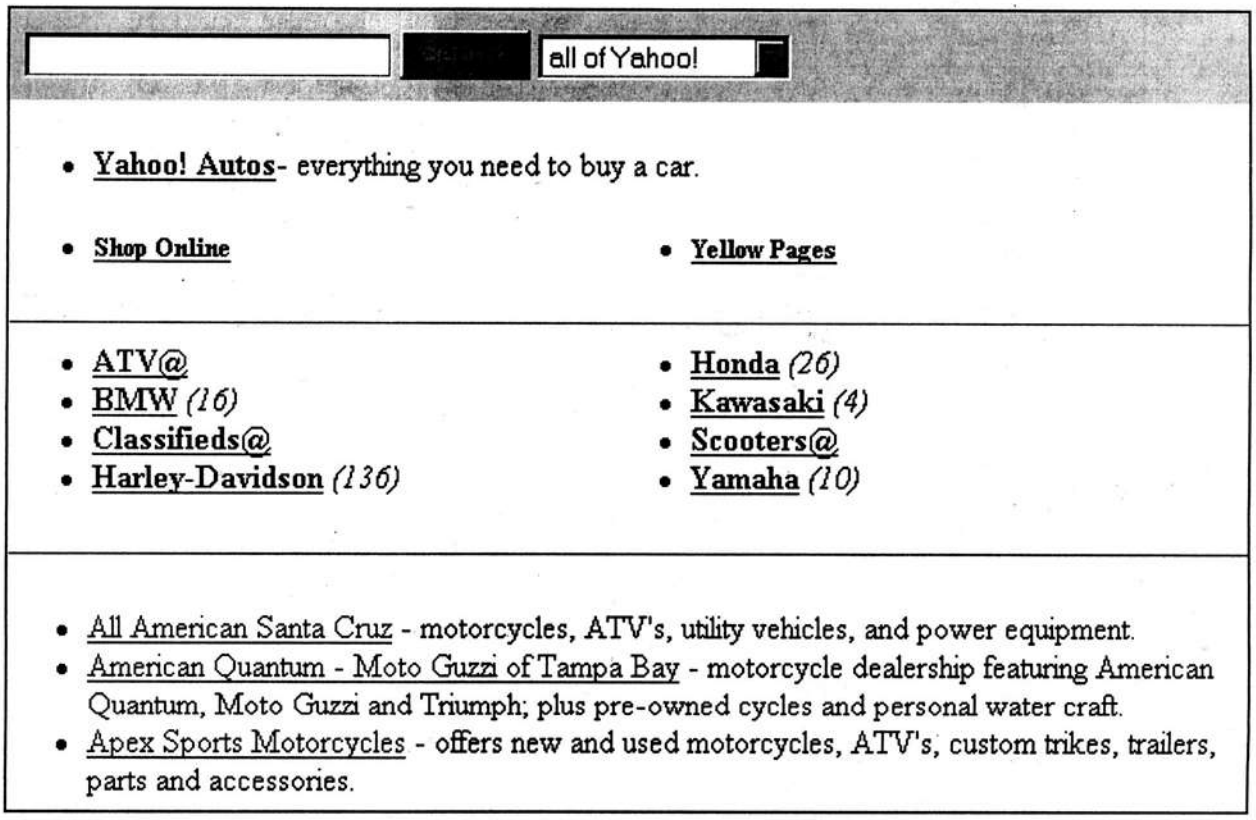
Kerek zárójelek között az osztályhoz tartozó találatok száma látható. Azokat az alosztályokat, amelyek alapvetően nem ide tartoznak, noha itt is feltüntették őket, a @ jelöli.

A helyzet azonban ennél bonyolultabb. A szerkesztők friss szemléletét minden jel szerint nyelvészeti szempontok sem kötik gúzsba; nem sokat foglalkoznak például a homonimák megkülönböztetésével. Gyakori, hogy ugyanazzal a névvel a rendszeren belül másik helyen másik osztályt is jelölnek, amelynek vagy nem ugyanaz a terjedelme (nem azonosak a hozzá besorolt információtételek), vagy nem ugyanaz a felosztása (nem azonosak az alatta megjelenő alosztályok). Például következőképpen kellene megkülönböztetni a többi, ugyanilyen nevű osztálytól: „Motorkerékpárok (a szabadidő és a barkácsolás szempontjából)". A szerkesztők nyilván abból indulnak ki, hogy maga a hierarchialánc is definiálja a jelentést. Hozzá kell azonban tenni, hogy „adott esetben". Más esetekben ugyanis eltérő hierarchialáncokban ugyanaz az osztály szerepel (pl. Motorkerékpárként), azaz az eltérő hierarchialánc nem definiál eltérően.

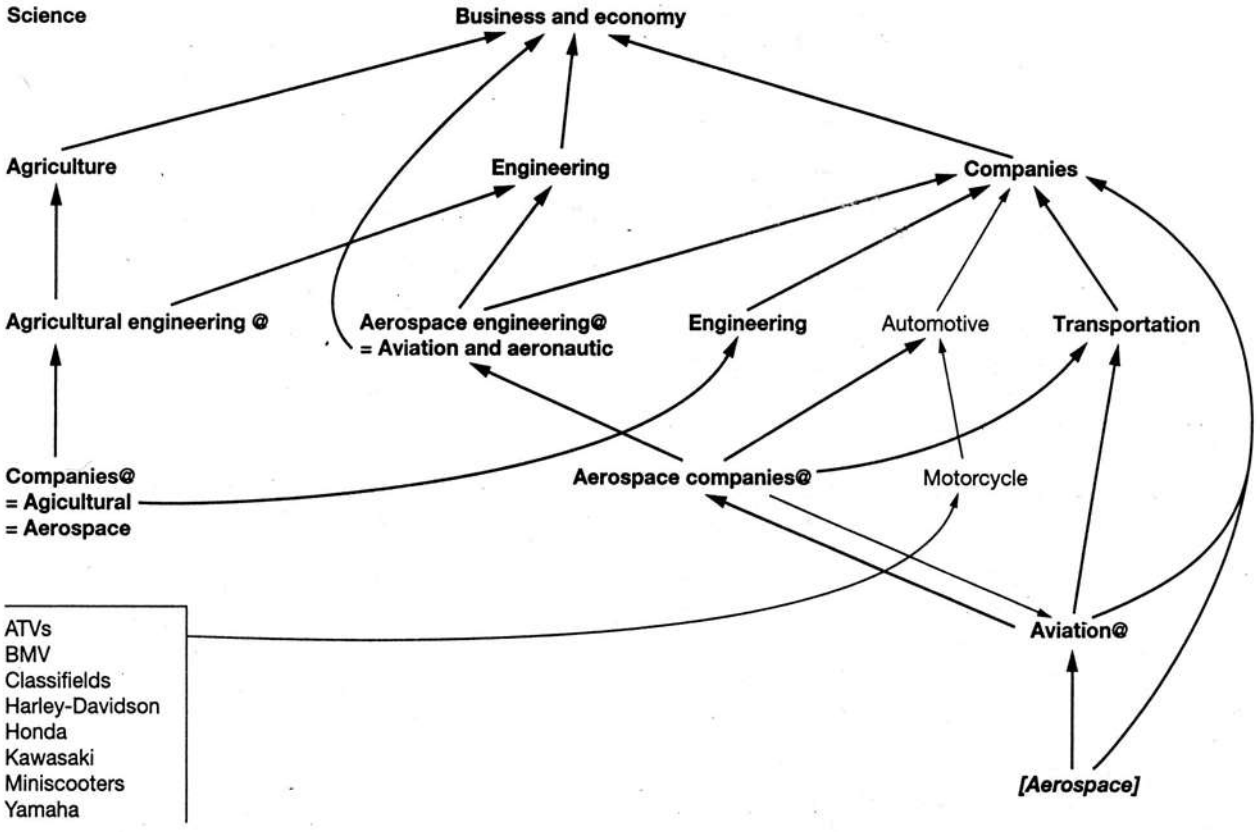
4.4.4.3 A struktúrák gazdagsága

Hogy ezeknek az osztályozási rendszereknek a rejtett szerkezeti bonyolultságát jobban lássuk, a *4. ábrán* a Yahoo! osztályozási rendszerének egy részletét kiemeltük, és címkézett irányított gráffal ábrázolva mutatjuk meg.

Az előbbiekben tárgyalt Motorkerékpárok osztály összefüggéseit a jobb elkülöníthetőség kedvéért nem félkövéren jelenítettük meg.



3. ábra **A Motorkerékpárok osztályának alosztályai és a Motorkerékpárok osztályba sorolt találatok jegyzékének eleje**



4. ábra **A Yahoo! polihierarchikus osztályozási rendszerének részlete címkézett, irányított gráf formájában**

A gráf alapján a következők ismerhetők fel.

Az Agricultural enginering (Agrotechnika) egyrészt az Agriculture (Mezőgazdaság), másrészt – @ jelöléssel – az Engineering (Mérnöki tudományok/Technika) alosztálya.

Az Aerospace engineering (Repüléstechnika) az Engineering és a Companies (Cégek), továbbá Aviation and aeronautic (Légügy/Repüléstan) néven a Science (Természettudomány) alosztálya, mely utóbbinak ugyanakkor tranzitív alárendeltje.

Az, hogy ugyanazt az osztályt más néven a tranzitív fölérendelt alá rendeljék, hajmeresztő a hagyományos osztályozási rendszerek ismerőjének (olyan ez, mintha a Kutyát egyrészt alárendelnék a Háziállatnak, ugyanakkor Eb éven az Állatnak, melynek ugyanakkor a Háziállat a közvetlen alárendeltje). A piaci viszonyok terén iskolázott rendszertervező viszont abból indulhatott ki, hogy a Természettudományok felől nézve jobban fest az általánosabban megfogalmazott osztálymegnevezés (Légügy...), nem pedig a Repüléstechnika, amely viszont a Technika felől nézve adekvátabb osztálynév.

Azt is észre kell venni, hogy az Aerospace enginering az Engineering alá rendelve valójában olyan osztályt képvisel, amely a repüléstechnikára vonatkozó *információk* tételeit tartalmazza, a Companies alá rendelve pedig azt, amely a repüléstechnikával foglalkozó *cégek* információit tartalmazza. Ennek a példának az esetében nincs a Yahoo!-ban különbség a két osztály terjedelme (információtételei) között.

Az Engineering esetében azonban van. Ebből ugyanis két osztályt találunk, de ez a két osztály nem ugyanaz: a Cégeknek alárendelt osztály ugyanis – melyet dőlt betűvel jelenítettünk meg – csak a műszaki tevékenységeket végző cégek információtételeit tartalmazza, a Természettudományoknak alárendelt Engineering ezzel szemben minden, a technikára és a műszaki tudományokra vonatkozó információtétel osztályozására való.

A dőlt betűvel megjelenített Engineering alárendeltje az Agricultural (Mezőgazdasági) [így, jelzősen], amely az agrotechnikai cégek információtételeit tartalmazza. Ugyanennek az osztálynak az Agrotechnika alárendeltségében viszont Companies (Cégek) a neve. Ha belegondolunk, ez egész logikus: az Agrotechnika felől nézve cégekről, a műszaki cégek felől nézve meg „mezőgazdaságiról", azaz Agrotechnikai (cégekről) van szó.

Talán a legmerészebb húzás, amikor ugyanazt az osztályt alárendelik egy másiknak, ugyanakkor fölérendelik neki. Ez a helyzet az Aerospace (= Aerospace companies @) és az Aviation között. De ha meggondoljuk, hogy ezekben az osztályozási rendszerekben egyáltalán nincs pontosan meghatározva, hogy mit is értünk tulajdonképpen azon a relación, amely az egyes osztályokat összekapcsolja, ez a megoldás korántsem olyan hajmeresztő, mint ahogy logikai szempontból látszik. Eddig ugyanis abból indultunk ki, hogy az internetkatalógusok osztályozási rendszerei hierarchikusak, és alapvetően csak alá-fölé rendeltségi kapcsolatokat tartalmaznak. Valójában azonban olyan rendezőrendszerekről van szó, amelyekben nincs egyértelműen definiálva a kapcsolat: lehet hierarchikus (az esetek többségében), de van, amikor egyszerűen csak annyit jelent, hogy „lásd még". Az Aerospace és az Aviation között valójában az utóbbi összefüggésről lehet szó, és ez logikailag teljesen megengedett. Más lapra tartozik, hogy ezekben az osztályozási rendszerekben a mindenkori definiálatlan relációt csak az jelöli, hogy „az egyik következik a másik után". Ha a tezauruszszabvány szerint pontosan jelölnénk a tárgyalt esetet, az *5. ábrán* látható szócikkeket kapnánk.

Transportation	Aerospace	Aviation
A Aerospace	F Transportation	F Transportation
Aviation	X Aviation	X Aerospace

5. ábra **Yahoo! összefüggések szabványos tezauruszcikk formában**

4.4.4.4 Az osztályozás

A HTML-dokumentumok tartalmi leírása egyrészt abból áll, hogy besorolják a megfelelő osztályba, és az osztály dokumentumhoz kapcsolt megnevezése vagy jelzete egyben „leírás" is. Ez a tartalmi leírás azonban formális adatok (szerző, cím, kiadó, annotáció stb.) nélkül használhatatlan, mert nincs, ami a dokumentumot egyértelműen azonosítaná (az URL kivételével).

Az internetkatalógusokban intellektuálisan dolgozzák föl a HTML-dokumentumokat, ezért nem készül keresőprogrammal („keresőgéppel") automatikusan a formális adatokról dokumentumleírás (lásd e tanulmány I. részében a 3. ábrát). A formális leírásokat tehát szintén manuálisan kell elkészíteni, hogy létrejöjjön a metaadatokat (szerző, cím stb.) tartalmazó teljesebb másodlagos információtétel. Ezeket az esetek jelentős részében maguk a beküldők, tehát laikusok készítik el.

Az önkéntesen beküldött tételek számos katalógusban többségben vannak, de jóformán minden kereskedelmi célú szolgáltatásban rendelkezésre állnak bejelentési űrlapok. A Yahoo!-ban pl. az egyes osztályok lapjának alján található „Suggest a site" (másutt „Add a site here", „Add URL" stb.) csatolóval hívható be. Bennük megtalálhatók a rovatok az osztályozás, a cím (Title), URL, tartalmi kivonat (Description) stb. számára.

„A beküldött űrlapok adatait elvileg a szerkesztők felülvizsgálják. A tapasztalatok arra utalnak, hogy ez annál nehezebb, mennél szabadabban adhatók meg az ada-

tok, s annál nagyobb munka az egységesítésük. Mivel a mennyiségi növekedés miatt egyre kevésbé képesek a szolgáltatások saját erőből elvégezni a leírásokat, a metaadatok megállapítását igyekeznek a beküldőkre bízni. Ennek érdekében részletező űrlapok szükségesek, hogy a laikus mindent jól értsen (jól példázzák ezt a Magellan's Reviews és az NISS űrlapjai). A metaadatok előrehaladt nemzetközi szabványosítása, különösen pedig a Dublin Core metaadatszabvány az internetkatalógusok információtételeiben a leírások egységesülését segíti elő. A fejlettebb katalógusokban, mint amilyenek a szakmai információs kapuszolgáltatok, részletesebb és színvonalasabb rekordleírási szabályzatok alakulnak ki.

A tételek megjelenítése és találati értékelése szempontjából különösen a tartalmi kivonatnak van nagy jelentősége. Számos katalógusban ez még csak egyetlen mondat. A részletesebb leírásokat szolgáltató katalógusokban a tartalmi kivonatot szemlének (review) is nevezik, de ezek sem lépik túl a hagyományos annotációk terjedelmét.

Különösen az igényesebb szolgáltatásokban fordul elő, hogy az osztályozási rendszer valamelyik osztályába besorolt dokumentumhoz még tárgyszavakat vagy deszkriptorokat lehet rendelni. Mivel számos internetkatalógusban nemcsak böngészni lehet az osztályozási rendszer hierarchikus szerkezete mentén, hanem természetes nyelven is le lehet kérdezni az állományt, a tárgyszavak és deszkriptorok kereshetőbbé teszik a HTML-rekordokat. A Beyond Bookmarks [1] szerint 1999 végén 20 szolgáltatásban használtak szabványosított természetes nyelven alapuló szótárt, ezen belül 13 tezauruszt. Az Engineering Electronic Library például hierarchikus osztályozási rendszere mellett saját tezauruszt is használ. A NetFirst a Kongresszusi Könyvtár osztályozási rendszerének (LCC) dokumentumtipológiája szerint is osztályoz.

Vannak internetkatalógusok, melyekben intellektuálisan értékelik a dokumentumokat (pl. Argus Clearinghouse, Lycos/Point Top 5%, Excite Reviews és Magellan's Reviews). Többnyire 1 és 5 közötti skála értékeit adják meg pontokban" [10].

4.4.5 Lekérdezés az internetkatalógusokban és a kereső- és böngészőszolgáltatás egyesítése

Általános jelenség, hogy az internetkatalógusokban nemcsak a hierarchikus osztályozási rendszerben lapozgatva lehet böngészni, hanem megadható külön ablakban természetes nyelven a keresett szó. Ha ez megegyezik a rendszer valamelyik osztályának nevével, vagy nevének részletével, akkor a kereső rögtön az adott osztálynál találja magát (így kérdeztük le pl. a 2. ábrán a „motorkerékpár" kifejezést a Yahoo!-ban).

E nem különösen szellemes segítségen kívül azonban megfigyelhető tendencia, hogy a katalógusokat integrálják az indexelőszolgáltatásokba. A katalógusok adatbázisainak mérete lényegesen kisebb, mint az indexelőszolgáltatásoké. Mivel többnyire intellektuálisan osztályoznak, a teljességre eleve nem törekedhetnek. Annak érdekében, hogy még több releváns adatot szolgáltassanak, hogy ők legyenek a „legjobbak a weben" („the Best of the Web"), „keresőgépet" is alkalmaznak, és az így megvalósítható lekérdezést szorosan vagy kevésbé szorosan összekapcsolják a böngészéssel. Általános gyakorlat, hogy az osztályozási rendszer bármelyik pontjából mind az osztályozási rendszer megnevezései, mind pedig a „keresőgép" által indexelt állomány lekérdezhetők. A szorosabb integrációra jellemző példa az Excite meg a Magellán, melyben kiválasztható, hogy az egész adatbázisban, a katalógus intellektuálisan feldolgozott és értékelt tételei (rated and reviewed sites) között, vagy a gyerekek számára is megengedhető „zöld" tételek állományában („green light sites") kívánunk keresni. A puszta egymás mellett létezésre is számos példa akad (mint a Lycos német változatában).

4.4.6 Regionális katalógusváltozatok

A nagyobb internetkatalógusok egyre több nemzeti/nyelvi változatot is létrehoznak. Ezek jelentős része valójában teljesen önálló, csak éppen átveszi a know-how-t. Bennük csak az adott ország, régió forrásait dolgozzák föl. A Yahoo! jelenleg már tucatnyi nemzeti változatban létezik, de a Lycos sem nagyon marad le mögötte. Az előbbiben a World Yahoo! osztály alatt találhatók meg az egyes nyelvi változatok, amelyek nem pontos másolatai az angolnak, hanem az adott ország körülményeihez alkalmazkodó fejlesztések (van már kínai nyelvű is).

A tendencia – kevésbé erőteljesen – az indexelőszolgáltatások terén is megfigyelhető, jellegzetes példa erre az AltaVista magyar változata, a Matáv AltaVizsla indexelőszolgáltatása, vagy a nemzetközi MetaCrawler, és annak német változata, a MetaGer.

Nem tévesztendők össze a nagyobb keresőszolgáltatók regionális változatai az önálló nemzeti jellegű keresőszolgáltatásokkal. A magyar Hungary.Network HUDIR internetkatalógusa például teljesen önálló fejlesztés, noha korai változatában a Yahoo! mintáját követte; az első magyar indexelőszolgáltatás, az ugyancsak a Hungary.Network által fenntartott Heuréka pedig az AltaVistától teljesen függetlenül jött létre.

4.5 Speciális adatbázisok

Mind az indexelőszolgáltatásokra, mind az internetkatalógusokra jellemző, hogy a keresőprogramokkal („keresőgépekkel") végzett lekérdezést, illetve a hierarchikus katalógusaikban végezhető böngészést különféle kisebb adatbázisokkal és szolgáltatással is kiegészítik, amelyek többsége

önálló, tágabb értelemben vett, nagyon specializált keresőszolgáltatásnak is tekinthető. Afféle miniatűr online szolgáltatókká válnak. A nagyobb piaci részesedés és a reklámbevétel növelésének reményében létrehozott kiegészítő adatbázisokra jellemző, hogy általános érdeklődésre tarthatnak számot, ingyenesek, és könnyen kezelhetők. Ezek az adatbázisok a hierarchikus rendszertől elkülönített osztályok (Bevásárlás, Apróhirdetések, Szótárak stb.) formájában jelennek meg a portállapokon. Az osztályozáselmélet szemszögéből felsoroló, enumeratív osztályozási rendszert alkotnak. (A Yahoo! esetében ilyen enumeratív rendszert képviselnek az 1. ábra felső részén a vízszintes vonal fölötti osztályok.) Könyvtárszervezési szempontból azt mondanánk, hogy ahány osztálytípus, annyiféle gyűjtőköri forrástípus.

Az osztályok (adatbázisok) típusai:

Szakterületek, tudományok, tevékenységi körök

Arts & Humanities (Művészet és társadalomtudomány)
Bussines & Economy (Kereskedelem és gazdaság)
Computers & Internet (Számítástechnika és internet)
Education (Oktatás-művelődés) stb.
Ezek az osztályok felelnek meg a dokumentumok hagyományos osztályozási rendszereiben alkalmazott osztályoknak, de itt is lépten-nyomon érheti az embert meglepetés: valamelyik szakterületen belül felbukkanhat apróhirdetéseket tartalmazó osztály, vagy tényadatokat tartalmazó osztály stb. (Az 1. ábrán a felső vízszintes elválasztó vonal alatti hierarchikus rész ezekből az osztályokból épül fel.)

Kereskedelmi jellegű osztályok

Shopping (Bevásárlás)
Travel Agent, Travel Finder (Utazási irodák), Book a hotel (Szállodafoglalás)
Buy a car, Buy a home (Autóvásárlás, Lakásvétel)
Classified (Apróhirdetések, üzleti)
Personals (Apróhirdetések, személyi)
Careers, Jobs (Álláshirdetések)
Ezek elsősorban arra valók, hogy az adásvételt támogassák. Az osztályok erősen válogatott, csak a rendelésfeladás szempontjából szóba jöhető szakterületek. Ezeken belül a besorolt információtételekből kiindulva megrendelhetők árucikkek, utazáshoz jegyek, elérhetők a hirdetések feladói.

Adattárak, címek, helyek osztályai

Community (Közérdekű és igazgatási információk)
Yellow Pages (Szakmai telefonkönyv), White Pages (Betűrendes telefonkönyvek)
People Search (Drótpostacím és személykeresés), WhoWhere (Ki kicsoda)
Search for Missing Children (Eltűnt gyerekek)
Books (Könyvek)
Auctions (Kiállítások, árverések)
Maps, City Guide, Roadmaps (Térképek)
Pictures & Sounds (Képek, Hangdokumentumok)
Photo Finder (Fényképek)
Videos, Video Search (Videofilmek)
Dictionares, thesauri (Szótárak, tezauruszok)
Free Software, Free Homepages (Ingyen beszerezhető programok)
Airlaine Tickets (Repülőjegyek), Menetrendek
Ezekben az osztályokban fehér és sárga telefonkönyvek, cégek, személyek adatait tartalmazó információtételek, egyéb céginformációk találhatók. Elmondható, hogy a segítségükkel az internethez már kapcsolódó országok túlnyomó részében szinte minden cím megtalálható. A térképek esetében helyek azonosíthatók vizuálisan. A szótárak, valamint a tezauruszok egy része többnyelvű.
Egyes keresőszolgáltatások felveszik a közlekedési vállalatok menetrendjeit is enumeratív osztályozási rendszerükbe.
Különlegesség – például az Infoseekben – a személyes honlapokat tartalmazó adatbázis.

Hírek, tényadatok

Today's news, What's News, What's Cool, Headlines (Aktuális hírek)
Stock Quotes (Tőzsdehírek)
Sports (Sporthírek)
Weather (Időjárás-jelentés)
TV (Tévéműsor)
Ezekben az osztályokban tényadatok szerepelnek. A híreket a nagyobb hírügynökségektől veszik át, olykor óránként aktualizálják őket.

Segítségek, gondűzők

Calendar (Naptár, események)
Horoscopes (Horoszkópok)
Games (Játékok)
Pager (Letöltő)
My Yahoo! (Testre szabható Yahoo!)
Yhooligans (Kapcsolatok)
E-mail (Drótposta-bejelentkezés)
Funny Site (Vicckereső)
Ezekben az osztályokban a mindennapokban hasznos eszközök és játékok találhatók. Többségük valójában nem is osztály (nem információtételeket tartalmaznak), hanem speciális szolgáltatások belépőpontjai.
Szolgáltatásként – a szótárakat és tezauruszokat kiegészítendő – feltűnnek az automatikus fordítórendszerek is; velük tetszés szerinti szöveg gépi fordítása végeztethető el a nagyobb világnyelvek között, az URL megadásával egész honlapok is lefordíthatók.*

4.6 Terminológia

Ha az internetkatalógusokban osztályozási rendszerek alapján végzett keresésről, azaz „szisztematikus lapozásról", vagy „strukturált gyűjteményekben való navigálásról" van szó, mindig böngészésről beszélünk. Az angol és német szakirodalomban túlnyomórészt „browsing" a neve.

Az internetes indexelőszolgáltatásokban természetes nyelvi kifejezésekkel, tárgyszavakkal, deszkriptorokkal és a Boole-műveletek segítségével végzett keresésre az általános keresés vagy a lekérdezés szót használjuk (searching, scanning, Suche).

* A leginkább elterjedt Systran fordítórendszert alkalmazza az AltaVista Translator (*http://babelfish.altavista.digital.com/*) és a Go translator service (*http://translator.go.com/*).

Ha a dokumentumok szövegén belül hipercsatolók felhasználásával – tehát nem szisztematikus rendszer mentén – kutakodunk, „szörfölésről" (surfing, Surfen) beszélünk. Az utóbbival összefüggésben beszélnek olyan keresésről, amelynek során értékes dolgok fedezhetők föl kevéssé valószínű helyeken is (serendipitous discovery); ezt nevezzük „innovatív vagy felfedező keresésnek". Ellentéte a hagyományos eszközökkel végzett böngészés és lekérdezés, amelyekre összefoglalóan angolul (a „tunnel vision" = csőlátás analógiájára) a nem túl hízelgő „tunneled searching" („kötött pályás keresés") kifejezést használják.

A böngészés, lekérdezés és szörfölés, illetve a kötött pályás és az innovatív keresés szakterülete az információkeresés (information retrieval). E szakterülethez tartozik az automatikus indexelés és osztályozás is.

Hagyományos körülmények között a szörfölésnek a könyv teljes szövegében végzett lapozás, a böngészésnek a tartalomjegyzékben, a lekérdezésnek a név- és tárgymutatóban végzett keresés felel meg.

5. Internetes dokumentumok és formátumok

5.1 A digitális és a virtuális dokumentum fogalma

Az internet különféle dokumentumai alkotják a virtuális könyvtár potenciális gyűjtőkörét. E gyűjtőkör dokumentumai túlnyomórészt nem kerülnek a könyvtár fizikai értelemben vett állományába, ezért könyvtári tárolási szempontból ezek a dokumentumok virtuálisak.

A digitális (csak digitális formában létező) és digitalizált (eredetileg nem elektronikus formában készült) dokumentumok a digitális könyvtár gyűjtőkörét alkotják. Ezek a dokumentumok lehetnek az internet HTML-dokumentumai, de olyanok is, amelyek fizikai értelemben is a könyvtár állományába tartoznak, tehát tárolási szempontból nem virtuálisan, hanem fizikailag léteznek (pl. CD-ROM-kiadványok). Az elektronikus könyvtár lényegében a digitális könyvtár szinonimája (egyes szakemberek szolgáltatási–működési szempontból elektronikus, feldolgozási–tárolási szempontból digitális könyvtárról beszélnek).

Az egyes könyvtárak által feldolgozott, de állományba nem vett HTML-dokumentumok az adott könyvtár szempontjából virtuálisak.

Tágabb értelemben virtuális minden olyan dokumentum, amely nem tartozik az adott könyvtár állományába, de a könyvtáron keresztül, annak másodlagos információi alapján mégis elérhető. A könyvtárban például tárolják a nem állományi dokumentum katalógustételét, amelyből az elsődleges dokumentum tárolási helye megállapítható. Szűkebb értelemben azok a HTML- és egyéb hálózati elektronikus dokumentumok virtuálisak, amelyek nem tartoznak a könyvtár állományába, de a könyvtár állományába tartozó másodlagos információk alapján távoli hozzáféréssel elérhetők. A távoli hozzáférésű elektronikus dokumentumok tehát mindig virtuális dokumentumok.

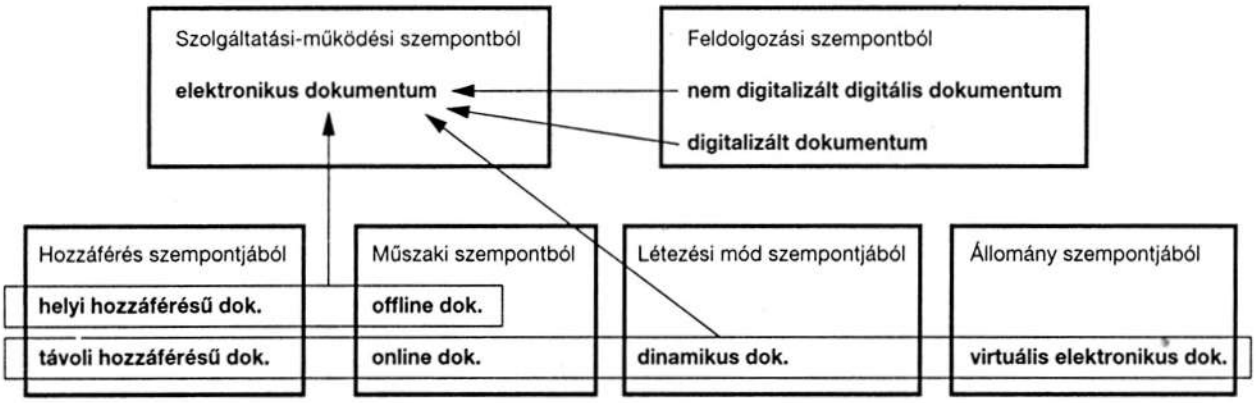
Mindezek alapján a szűkebb értelemben vett virtuális könyvtár a digitális könyvtár egyik fajtája (másik fajtája pl. a CD-ROM könyvtár). Fordítva ez nem igaz: nem minden digitális könyvtár virtuális.

Digitális, de elsősorban virtuális könyvtári környezetben a dokumentum fogalma problematikussá válik, ezért inkább digitális objektumokról beszélnek (az ilyen típusú dokumentumok meghatározásának kérdésével részletesen foglalkozik [4] és [9]). Ezek megfelelnek a hagyományos könyvtárak állományi egységeinek (könyvek, időszaki kiadványok, térképek, zeneművek stb.). Mind a digitális és digitalizált, mind a hagyományos könyvtári dokumentumok elsődleges adatokat tartalmaznak, és maguk is elsődleges dokumentumok. Beszélnek még offline és online elektronikus dokumentumokról. Az előbbiek az adott könyvtár állományában vannak (pl. CD-ROM típusú dokumentumok), az utóbbiakat csak külső online hozzáféréssel lehet használni. Az offline dokumentumok a helyi hozzáférésű elektronikus dokumentumok, az online elektronikus dokumentumok pedig a távoli hozzáférésűek. Az utóbbiak felelnek meg a virtuális elektronikus dokumentumoknak (nevezik ezeket dinamikus dokumentumoknak is). A terminológiát a *6. ábrán* címkézett, irányított gráffal szemléltetjük. A közös vastag vonallal keretbe foglalt kifejezések közös szempontból megfogalmazott megnevezések. A közös vékony keretbe foglalt kifejezések egymás szinonimái. Ez könnyen ellenőrizhető: ha pl. minden „elektronikus dokumentum" „digitális dokumentum", és minden „digitális dokumentum" „elektronikus dokumentum", akkor a két megnevezés ugyanazt a dokumentumot jelöli, tehát szinonim.

Az elektronikus (digitális/digitalizált és virtuális) dokumentumok és a hagyományos dokumentumok között az alapvető különbség, hogy az előbbieknek mind a tárolása, mind az olvashatósága ugyanabban a gépi keretrendszerben játszódik le. (A hagyományos dokumentumokat nem gép tárolja, noha géppel [be/le]olvashatók.) A digitálisan feldolgozott dokumentumot a számítógép mintegy „belülről" ismeri, azaz minden adatához funkcionálisan hozzáfér. Ebből következik, hogy az elektronikus dokumentumszövegek gépi kezelési szerkezetének funkcionális szempontú szintaktikai–sze-

mantikai egységesítése közérdek: ilyen módon válik ugyanis lehetővé, hogy a dokumentumokat (objektumokat) a legkülönfélébb információs szervezetek nehézségek nélkül kezelni tudják, amikor arról van szó, hogy szolgáltatni kell őket.

tumoké, szintaktikai és szűkebb értelemben szemantikai szabályokat biztosít a szöveg hierarchikusan rendeződő elemeinek formális leírásához. Alapvető különbség a MARC formátumokhoz képest, hogy az SGML segítségével ugyanazt a do-



6. ábra **Az elektronikus dokumentumok átfogó tipológiája**
A digitális dokumentumok egyik fajtája az eleve digitálisan készült („nem digitalizált digitális") dokumentum, és a digitalizált dokumentum

5.2 Formátumok

5.2.1 Elsődleges dokumentumok formátumai

Ebből a célból születtek meg az elektronikus dokumentumok formátumszabványai, amelyek alapján a digitális/digitalizált szöveg bizonyos szerkezeti egységei egységesen kódolhatók (minősíthetők). Rendeltetésüket tekintve nagyon hasonlóak azokhoz az adatcsere-formátumokhoz, amelyeket a másodlagos adatokra vonatkozó dokumentációs és könyvtári adatok számára alakítottak ki jóval korábban. A különbség, hogy elektronikus dokumentumok esetében a szabványosítás a *közvetlen számítógépes* kezelhetőség és olvashatóság következtében már az elsődleges dokumentumra vonatkozóan megvalósítható. Mivel a nyomtatott dokumentumok ma már számítógépek igénybevételével készülnek, létezik elektronikus változatuk, amelyek előbb-utóbb bekerülnek a tárolandó és kereshető állományok világába.

Az elsődleges elektronikus dokumentumok szerkezetét az elsődleges dokumentumon belül leíró metaadatszabvány az 1986-ban elfogadott (ISO 8879) SGML (Standardized General Markup Language = Szabványos Általánosított Jelölőnyelv). Készítői az egyszerűbb és a tényeknek megfelelőbb „formátum" vagy szabvány helyett a „nyelv" megnevezést használták, noha nincs szó olyan értelemben mesterséges nyelvről, mint amilyenek a programnyelvek (hiszen a formátum, akárcsak az űrlap vagy a könyv, nem nyelv, hanem valamilyen nyelven kifejezett információ, adat, esetünkben szabvány). Az SGML-szabvány elsődleges feladata ugyanaz, mint a MARC formákumentumot különféle – konkurens – szerkezetekben is le lehet írni. Az adott, ténylegesen használt leírás neve Document Type Definition (DTD) [13].

A HTML (Hypertext Markup Language = Hipertext Jelölő Nyelv) [6] a web közismert adatformátuma, valójában SGML-alkalmazás, vagyis egy lehetséges DTD, amelyet a World Wide Web Consortium (W3C) definiált. A webnézegetők valójában olyan SGML-olvasók, amelyek csak egyetlen – viszonylag egyszerű – DTD feldolgozására alkalmasak. A HTML DTD elsősorban olyan alkotóelemeket tartalmaz, amelyek a képernyő-megjelenítést szabályozzák, vagyis minimális mértékben határozza csak meg az adat logikai–szemantikai szerkezetét, hierarchiáját. Mint ilyen, kevéssé alkalmas a jól visszakereshető, strukturált digitális objektumok rögzítésére. A kliens–szerver szerkezetű dinamikus keresőszolgáltatások megjelenése fokozatosan megváltoztatja ezt a helyzetet, melyről *Lou Burnerd*, az SGML szintaxison alapuló szemantikai rendszer, a TEI (Text Encoding Initiative) egyik szerkesztője így ír:

„Mégis, miért használjuk a HTML-t? A gazdasági, politikai és szociológiai érvek mellett van még egy eddig figyelmen kívül hagyott szempont: a web tartalmának jelentős része eredendően tiszavirág-életű. Ezek az anyagok csak itt és most kívánnak hatni, például terméket eladni, vagy egyszerűen szenzációt kelteni. Ebből következően semmi értelme ezekre több energiát pazarolni, mint a hasonló papírbrosúrákra. A gondot inkább az okozza, hogy éppen úgy a HTML-t kell használunk, ha fontos kézikönyvet digitalizálunk, mint ha éppen üdítőitalt reklámoznánk.

Valójában azonban még az értékesebb művek rögzítésénél is csak akkor tűnik föl a HTML gyengesége, ha a szerző vagy a kiadó szempontjából vizsgáljuk a hely-

zetet. Ha a képernyőkép tetszetős, az olvasó számára végső soron mindegy, hogy a korszerű objektumorientált adatbázis-kezelőből, postscript fájlból, vagy pedig feketemágiával előállított HTML-fájlból származik-e... A HTML-nek mint szerveroldali formátumnak van néhány nyilvánvaló hátránya. Noha a kezdeti költségek kicsik, HTML-dokumentumokkal aligha tanácsos komolyabb, hosszabb távú szolgáltatást indítani. A hivatkozások konzisztenciájának megőrzése már viszonylag dinamikus állomány esetében is rendkívül sok fejfájást okozhat" [2].

A megoldást minden jel szerint a tényleges SGML és a kurrens HTML-változat ötvözése jelenti, mindegyiket arra használva, amire való: valódi SGML formátumot használni a szerveroldalon, és HTML-t a kliensoldali megjelenítéshez. A gyors fejlődés jele, hogy a World Wide Web Consortium 1998 február elején adta közre az XML (Extensible Markup Language = Kiterjeszthető Jelölőnyelv) webszabvány első változatát, amely az SGML lényegesen egyszerűsített változata, többféle dokumentumtípus rögzítéséhez használható szabvány, szemben a régi HTML-lel, amely csak egyféle dokumentumtípushoz használható, s ezért a multimédiás környezetben is megállja a helyét [15]. (Számos, a kérdéssel összefüggő testületi dokumentum található az OMIKK Virtuális Könyvtárának oldalain [11].)

Mivel elvileg nincs akadálya annak (csupán megfelelő konvertálóprogramok kérdése), hogy a HTML és az XML formátumon belül a dokumentum típusát meghatározó leírást (ez a DTD nevű rész) a MARC formátumot használók áttegyék a saját formátumukba, csak idő kérdése, hogy az elektronikus dokumentumokat a könyvtárak automatikusan is átvegyék, és a saját igényeik szerint kezeljék. Az elektronikusdokumentum-formátumok kialakulása utal arra az ismeretelméleti felismerésre, hogy az internet (és dokumentumainak) megjelenésével csak ugyanaz fejlődik tovább, ami az írott történelem kezdetén a könyvtárakkal elkezdődött.

5.2.2 Másodlagos adatok formátuma (metaadat-formátum)

Az elsődleges dokumentumokra vonatkozó adatok a másodlagos adatok. Ilyenek a bibliográfiai leírás szabványosított adatelmei, továbbá minden, a dokumentumok tartalmi leírására felhasznált információkereső nyelvi/osztályozási adat (kulcsszó, tárgyszó, deszkriptor, osztályozási jelzet). Digitális könyvtári környezetben ezeket az adatokat többek között metaadatoknak nevezik, ilyen adatokat határoznak meg az előbb ismertetett formátumszabványok. Segítségükkel az elsődleges elektronikus dokumentumok egységes gépi kezelése valósítható meg.

Metaadat tehát szűkebb értelemben az internetforrások intellektuálisan vagy automatikusan létrehozott másodlagos adata, melyet vagy magába az elsődleges dokumentumba ágyaznak be, vagy csatolókkal kapcsolnak hozzá. Korántsem olyan nagy a választékuk, mint a bibliográfiai formátumokban rögzített adatelemeké, és nem olyan komplexek, mint az utóbbiak.

Szükségesnek bizonyult maguknak a metaadatoknak az egységes elektronikus kezelése is. Ide tartozik a metadatoknak az elsődleges dokumentumokból (digitális objektumokból) való kinyerése vagy kiszámítása, a dokumentumok számítógépes leírása. Ezek az adatok a funkcionálisan strukturált (pl. SGML) dokumentumok esetében rendkívül könnyen kinyerhetők, noha erre alapul szolgálhat az elektronikus dokumentum teljes szövege is. A sokféle metaadat-formátum léte hívta életre a Dublin Core (DC; dublini alap[mag]-metaadatok) formátumát, amelynek 1999. 09. 09-i 1.1 változata 15 metaadatelemet tartalmaz az elektronikus dokumentumok egységes leírására (és tegyük hozzá: eme adatelemekből felépülő rekordok cseréjére is) [3]. Ez a viszonylag egyszerű formátum független attól a szintaxistól, amelyben az elektronikus dokumentumot funkcionálisan strukturálták (elvileg tehát alkalmazható nemcsak SGML-dokumentumokra is). Minden adatelemnek több értéke lehet (ismételhető) és opcionális.

A DC metaadatelemei az elektronikus dokumentumok katalogizálását teszik lehetővé. Közöttük van a „Tárgy" (<Subject>) azonosítójú metaadatelem, amelynek ismételhető értékei kulcsszavak, tárgyszavak, deszkriptorok, osztályozási jelzetek lehetnek.

Mivel szükség van a DC formátumot kiegészítő információkra is (pl. a felhasználás feltételeire), született erre vonatkozó átfogó ajánlás (architektúra, container architecture), amelyet Warwicki forrásleíró keretmegállapodásnak (Warwick Framework, Resource Description Framework) neveznek [11].

A fejlődés iránya, hogy a HTML-rekordok valamilyen formátum szerint egységesüljenek. A fejlődés a DC formátum irányába mutat.

A metaadat-szabványosítás terén két irányzat küzdelme figyelhető meg: a minimalisták szemében csak az a fontos, hogy a keresést megkönnyítsék (ezért legyen a lehető legegyszerűbb a formátum); a strukturalisták fontosnak tartják, hogy a digitális dokumentumnak legyen valamilyen azonosító jellegű, a bibliográfiainak megfelelő leírása is, hogy adatcsere esetén tudni lehessen, miről is van szó a tételek esetén.

A DC elsősorban a web számára kialakított szabványos formátum. A digitalizált (tehát eredetileg nem digitális) dokumentumokra nem alkalmazható kifogástalanul. A keresés szempontjából például a „Dátum" és a „Kiadó" adatelemek okoznak problémát, melyek a szabvány szerint nem az eredeti mű, hanem a digitalizált dokumentum adatai. Márpedig képzőművészeti alkotás vagy szépirodalmi mű esetében az eredeti mű dátuma és kiadója sokkal fontosabb, semmint hogy elhagyható lenne. Bibliográfiai szempontból a „Cím" is rendkívül problematikus, melyre semmiféle egységesítést nem írnak elő.

Irodalom

[1] Beyond Bookmarks: Schemes for organizing the web. *<http://public.iastate.edu/~CYBERSTACKS/CTW.htm>*

[2] BURNERD, L.: *<http://info.ox.ac.uk/ctitext/publish/comtxt/ct15/burnard.html>*

[3] Dublin Core Metadata Element Set. Version 1.1. Reference Description. Recommendation. *<http://purl.org/dc/about/element-set.htm>*

[4] GOLDEN D.–TÓTH T.–TURI L.: Virtuális örökkévalóság: objektumok a digitális könyvtárban. = Tudományos és Műszaki Tájékoztatás, 41. köt. 8–9. sz. 1998. p. 299–314. *<http://www.neumann-haz.hu/digitalis/studies/object/objects.html>*

[5] GŐZ Á.: Az Interneten elérhető információforrások katalogizálása. = Tudományos és Műszaki Tájékoztatás, 41. köt. 8–9. sz. 1998. p. 315–330. *<http://www.neumann-haz.hu/digitalis/studies/intercat/index.htm>*

[6] HTML (Hypertext Markup Language)

[7] KOCH, T.: Nutzung von Klassifikationssystemen zur verbesserten Beschreibung, Organisation und Suche von Internet Ressourcen. = Buch und Bibliothek, 50. köt. 5. sz. 1998. p. 326–335. *<http://www.ub2.lu.se/tk/publ/bubmanus.html>*

[8] KOCH, T.–DAY, M.: The role of classification schemes in Internet resource description and discovery. = EU project DESIRE. Deliverable D3.2.3. 1997. *<http://www.ub2.lu.se/metadata/subject-help.HTML>*

[9] KOLTAY T.–HORVÁTH P.: Digitális könyvtárak a világban. = Tudományos és Műszaki Tájékoztatás, 45. köt. 7. sz. 1998. p. 255–264. Bővebben: Digitális könyvtárak és projektek. Tanulmány. Neumann Ház, 1998. február. *<http://www.neumann-haz.hu/digital/studies/digital/digital.htm>*

[10] OHLER, A.: Browsingdienste im Internet. Berlin, Freie Universität, 1996. *<http://userpage.fu-berlin.de/~angele/bond/brows04.htm>*

[11] Az OMIKK Virtuális Könyvtára. Szerk. Válas Gy., Horváth P. 1999. 08. 16. *<http://www.omikk.hu/omikk/virkonyv/inet.htm>*

[12] HAKALA, J.–HUSBY, A.–KOCH, T.: Warwick Framework and Dublin Core Set provide a comprehensive infrastucture for network and resource description. = Report from Metadata Workshop II., Warwick, UK, April 1–3, 1996. *<http://www.ub2.lu.se/tk/dcwsrept.html>*

[13] SGML (Standardized General Markup Language). *<http://www.sil.org/sgml/sgml.html>*

[14] STEINBERG, S. G.: Seek and ye shall find (maybe). = Wired, 4. köt. 5. sz. 1996. p. 108–114., 172–182.

[15] XML (Extensible Markup Language). *<http://www.sil.org/sgml/sgml.html>*

Hivatkozott keresőszolgálatok*

AltaVista. AltaVista Inc. *< http://www.altavista.com>*

The Argus Clearinghouse. *<http://www.clearinghouse.net/docsy>*

Ariadne. *<http://ariadne.inf.fu-berlin.de:8000>*

BUBL Link (Bulletin Board for Libraries). Information Service. BUBL WWW Subject Tree – arranged by Universal Decimal Classification. 1996. *<http://www.bubl.ac.uk/link>*

EELS (Engineering Electronic Library). *<http://www.ub2.lu.se/eel/eelhome.html>*

Excite. Review *<http://www.excite.com>*

GERHARD (German Harvest Automated Retrieval and Directory). BIS Oldenburg *<http://www.gerhard.de>*

HUDIR. Budapest, Hungary.Network, 1996. *<http://www.net.hu/search>*

Kincskereső. Budapest, Elender Kft. 1999. *<http://eol.hu>*

Lycos. Point Top 5% *<http://point.lycos.com/categories/index.html>*

Magellan Review. The McKinley Internet Directory. *<http://www.mckinley.com>*

NetFirst. OCLC *<http://www.oclc.org/oclc/netfirst/faq.htm*, illetve *<http://www.ref.oclc.org.2000>*

NISS Information Gateway. *<http://www.niss.ac.uk/subject/index.html>*

Scorpion. *<http://purl.oclc.org/scorpion>*

Thesaurus compendium. *<http://www.darmstadt.gmd.de/~lutes/thesauri.html>*

Webcrawler. Select *<http://www.gnn.com/gnn/wic/support/about.rescat.html>*

WWW Virtual Library. *<http://vlib.stanford.edu/overview.html>*

Yahoo!. Yahoo! Inc. Search *<http://www.yahoo.com>*

---

* Ha indexelőszolgáltatással is rendelkező internetkatalógusokról van szó, a „Review" kiegészítő különbözteti meg a „Search" kiegészítővel jelölt indexelő válozattól (pl. Magellan Review).

Beérkezett: 1999. IX. 27-én.