

---

Vellucci, Sherry L.: *Herding cats: options for organizing electronic resources*. (Internet Reference Services Quarterly, 1. (4). 1996, 9-30.p.) c. tanulmányát **Mándy Gábor** tömörítette.

---

## Macskák terelgetése, avagy az elektronikus források szervezésének útjai

Sherry L. Vellucci

A szerző összehasonlítja azokat az eltérő módszereket, amelyeket a könyvtárosok, a kutatók, vagy a számítástechnikai szakemberek fejlesztettek ki az Interneten megtalálható elektronikus források szervezésére és elérésére. Megállapítja, hogy mindegyik csoport a maga képére szeretné alakítani a szervezési eszközöket (a könyvtárosok egy szuperkatalógust képzelnek el, a kutatók kódolt szövegben gondolkodnak, a számítástechnikusok pedig robot indexeket építenének), s hogy eddig az Internet-források fölötti bibliográfiai ellenőrzés olyan reménytelennek bizonyult, mintha macskákat akarnánk terelgetni.

Fontos különbség a könyvtár és az Internet között, hogy míg az előbbiben rögzített információtartalmakat módszeresen, egységes vezetés alatt szerveztek gyűjteménybe, addig az utóbbi jobban tükrözi a társadalom egészét átjáró információ „apály-dagály” jellegét, amely nem rendelkezik módszeresen kialakított struktúrával, nem stabil, nem szolgál ki egyetlen jól körülhatárolható használói csoportot, nincsenek látható céljai, és egyetlen intézmény sem gyakorol ellenőrzést fölötte. Ebből következik, hogy megszervezésére sincs felhatalmazva semelyik szakmai csoport.

Schwartz modellje a tudományos információ-cserére, a „lazán összekapcsolt rendszer” (loosely-coupled system) alkalmasnak látszik az Internet „rendszerének” kialakítására, azáltal, hogy mindegyik szakmai csoportot felelősnek, de egyben autonómnak tekintjük. Arra lehet számítani, hogy mindegyik szakmai csoport továbbra is olyan elérési módokat fejleszt ki, amelyek a saját elképzeléseik szerint a legmegfelelőbbek, s a legtöbb, amit remélhetünk, egy, a leírás számára alkalmas adatminimum közös értelmezésére irányuló kooperatív erőfeszítés. Mindegyik szakmai csoport ismereteit, tapasztalatait és eredményeit el kell ismerni, s olyan különféle bibliográfiai módszereket kell kidolgozni, amelyek egymásra rétegezhetők, csereszabatosak és egymásba konvertálhatók, egy lazán összekapcsolt szervezési rendszer keretében.

A könyvtári gyűjtemény olyan dokumentumok összessége, amelyeket az egymáshoz való kapcsolódásuk, valamint minőségük alapján választottak ki, s a használók számítanak erre a „hozzáadott értékre” az információáradattal való megbirkózásban. Anyagi okokból sok könyvtár változtatott állománygyarapítási politikáján, a dokumentumok birtoklása helyett a hozzáférhetőség megteremtésére helyezte a hangsúlyt, s ez nemcsak bizonytalanabbá tette a gyűjtemény fogalmát, hanem sokan a gyűjteményszervezésből adódó előnyöket is szem elől tévesztették. Alapvető fontosságú és kívánatos, hogy a gyűjteményi elv a könyvtár által nem birtokolt dokumentumok vonatkozásában is érvényesüljön. Így, bár a gyűjtemény kiterjesztett fogalmába belefér, hogy a helyi online katalógusban olyan művek is szerepeljenek, amelyek távoli adatbázisokból férhetők hozzá, megfelelő válogatási kritériumokat kell kidolgoznunk, illetve alkalmaznunk annak eldöntéséhez, hogy az Internetről milliószám elérhető elektronikus dokumentumok közül melyeket reprezentálunk a helyi online katalógusban.

Nem vehetünk fel a katalógusunkba minden Internet-forrást csak azért, mert hozzáférhető, de ugyanilyen rövidlátó politika lenne az online katalógust a helyi gyűjteményre szűkíteni, s minden egyebet a Web-keresési eszközökre bízni. Figyelembe véve az Internet-publikációk ellenőrizetlenségét, lektorátatlanságát, még inkább szükségesnek látszik, hogy a könyvtár gondos szelekció után felvegye katalógusába azokat a forrásokat, amelyeket a kutatók számára relevánsnak ítél. Annál inkább, mivel a Web-indexelő és visszakereső programok notóriusan pontatlanok, a használónak keresztül kell rágnia magát az irreleváns tételek tömegén, hogy egy-egy megfelelőt találjon. Ha egy minden releváns forrást magában foglaló katalógust nem is tudunk létrehozni, az lehetséges, hogy az olvasó egy egységes használói interfészt kezelve, ugyanarról a terminálról számos bibliográfiai eszközt használhasson. Olyan kifinomult „rétegezett” rendszeren dolgozunk, amely – közös parancsok és többnyelvű teauruszok segítségével – lehetővé tenné az online katalógus és különböző egyéb adatbázis egyetlen keresési folyamatban való használatát, azt a látszatot keltve az olvasóban, hogy mindvégig ugyanazt az eszközt használja. Ez a rétegezési technika, az adatok cseréje és konvertálása segít

elkerülni egy gigantikus integrált katalógus létrehozásával járó idő-, költség- és jogi problémákat. Mindez példa a „lazán összekapcsolt rendszerre”, és modellül szolgálhat az Internet megszervezésével kapcsolatos álmunkhoz.

Az Internet-források szervezésének első szintje a helyi online katalógus (OPAC), amely értékeli a dokumentumok használhatóságát, megadja leőhelyüket, szabványos leírásokat és hozzáférési pontokat nyújt (pl. a Kongresszusi Könyvtár tárgyszójegyzékének ellenőrzött szókészlete alapján). Ezeket a bibliográfiai adatokat a MARC formátumba kódolva áll össze az online katalógus. A katalogizálók körében számos próbálkozás történt a MARC rekordformátum alkalmassá tételére a távoli elektronikus dokumentumok leírására. Létrehozták a 856-os mezőt, amelynek almezői az elektronikus objektum leőhelyére és a hozzáférés módjára vonatkozó információkat tartalmaznak: a gazdaszámítógép neve (host name), címe (network address), a fájl neve és hozzáférési útvonala (path), hozzáférési instrukciók, bejelentkezési (logon) adatok, a tájékoztatás céljából felhívható személy (contact person), az URL stb. Magát az AACR2 szabványt is alkalmassá tették az elektronikus dokumentumok leírására (elsősorban a megjegyzések adatcsoportját felhasználva). Az OCLC 200 résztvevővel országos kísérlet keretében dolgozza fel az elektronikus forrásokat, s ennek keretében számos fontos kérdést vitattak meg. Az AACR2/MARC-hagyományok alkalmasak a továbbfejlesztésre, de számos nehézség adódik: a szabvány az egyedi, lineárisan hozzáférhető dokumentumok leírására született, nehéz alkalmazni a többszintű hiperszövegű objektumokra. Gondosan kiválogatott adatokra épít, amelyek analizálása és kódolása időigényes. Az elektronikus források leírásának egyszerűsítése ez ellen a hagyomány ellen hatna, s mivel a könyvtárvezetők pénzügyi megfontolásból külső forrásból származó bibliográfiai rekordokat próbálnak felhasználni, a bibliográfiai rekordok készítésének új módozatait kell keresnünk.

Egy ilyen új módozatot a társadalomtudományok számítástechnikai szakemberei találtak, amikor arra törekednek, hogy a katalogizáláshoz szükséges adatokat magában az elektronikus dokumentumban kódolják. Ez a TEI (Text Encoding Initiative), amely az SGML (Standard General Mark-up Language) segítségével állít elő géppel

olvasható dokumentumot, s amely egy rekordfejen jelenik meg, még a szövegtest előtt. A TEI-fej négy fő részből áll: a fájl leírásából, a kódolás leírásából, a szövegprofilból, valamint a szöveg módosításainak dátumaiból. Ez egyaránt szoros a könyvtárosnak (a katalogizáláshoz), a kutatónak (a dokumentum értelmezéséhez) és a számítógépes szakembernek (a dokumentummal végzett műveletekhez). A fej fájl-leíró része a nyomtatott dokumentum címlapjának felel meg, s ez a fej egyetlen kötelező része. A dokumentum létrehozóját különféle szoftverek segítik a kódolásban. A könyvtári szakemberek most fontolgatják a TEI-fej felhasználhatóságát a MARC rekord automatikus előállításában, s már léteznek olyan számítógépes programok, amelyek ezeket a TEI-fejeket MARC-formátumú bibliográfiai rekorddá konvertálják – pl. a Virginiai Egyetemen kifejlesztett TEI2MARC [TEI-ből MARC-ba]. Lehet, hogy a jelenlegi, külső adatbázisból származó rekordok másolására építő helyi katalogizálást (copy cataloging) ez a konverziós módszer váltja fel.

A második szint egy Internet központi katalógus lehet, amelynek egyik példája az OCLC InterCat Catalog nevű vállalkozása. Ez olyan MARC-rekordokból áll, amelyet több könyvtár katalogizálói hoztak létre, Web-interfészes kereséssel. Az adatbázis automatikusan bővül az OCLC-nek a 856-os mező kitöltésével jelentett tételekből. Az ebben a mezőben levő URL közvetlenül az adott dokumentumra mutat (hotlinked). Bár ez az Internet központi katalógus egy szinttel magasabban helyezkedik el, mint a helyi OPAC, annak sok hasznos jellemzőjét megőrzi: a dokumentumot könyvtáros választotta ki, a bibliográfiai leírás a dokumentum azonosítására, besorolására, értékelésére és megtalálására alkalmas adatokat tartalmaz, a hozzáférési pontok egységesítettek.

A feltárára váró dokumentumok száma az Interneten azonban olyan hatalmas, hogy ezek az egyedi eljárások már alkalmatlanok a feladat megoldására, új módszereket kell keresni. Az Internet-dokumentumokhoz való hozzáférésnek három alapvető módszere létezik: a közvetlen címek begépelése, a tallózó listák és a robot indexek.

Az első magától értetődő, nem igényel szervezési eszközöket, viszont csak akkor használható, ha tudjuk a dokumentum pontos leelőhelyét és a hozzáférési utat.

A tallózó listák (böngészők, browsing lists) a forrásokat tematikusan rendezik. A legáltalánosabban használt listák (pl. Yahoo) betűrendben sorolják fel a tárgyköröket, verbális témaköri hierarchiával. A tallózás egyszerű, de hátránya, hogy ismerni kell a témakörök szerkezetét, s kevés a keresztutalás. A listák egy-kettőre felduzzadnak, s az áttekintés még nehezebbé válik. A Yahoo (és néhány más lista) kulcsszavas keresést is lehetővé tesz. Más tallózó eszközök a könyvtárakon kívül jöttek létre, de hatással volt rájuk a könyvtári gyakorlat: a „Patrick’s Subject Catalog” és a „Mundie’s Cyber-Dewey” egyaránt osztályozási jelzeteket használ, amelyet az olvasók többsége megszokott, viszont keresztutalásokra és a tárgykör verbális pontosítására itt is szükség van, s a tárgyköri listák is nagyon felduzzadhatnak, ormótlanná válhatnak. A tallózó listák is munkaigényesek (a gyűjtés, rendezés, HTML-kódolás, a az esetleges annotációk miatt), amely jelentős hátrány, ha az Internet-források mennyiségére gondolunk. Amellett specifikus tételeket igen körülmenyes tárgykör szerint megtalálni.

A robot indexek automatikusan gyűjtik össze és indexelik az elektronikus forrásokat (főleg Gopherekből és FTP-helyekről), s a hozzáférés szavak, illetve kereső kifejezések szerint történik. A legnépszerűbb ilyen eszközök: Lycos, InfoSeek, WebCrawler és Alta Vista. Adatbázisaik rekordok millióit indexelik, de módszereik eltérőek, s ez igen eltérő keresési eredményekhez vezetnek. Legnagyobb hiányosságuk a hamis találatok nagy száma. (A szerző példának a Black beauty c. elektronikus dokumentumot használja; az Alta Vista robot index eredetileg 3400 találatot mutatott, s a kifinomult logikai operátorok használatával is mindössze 101-re sikerült leszűkíteni a találatok listáját.) A zaj oka, hogy ezek a robot indexek nem gondosan kiválogatott adatcsoportokkal dolgoznak, hanem az egész szöveggel, s még az utalásokat is találatként értelmezik. (Az egyik téves találat az Augsburgi hitvallás volt, mivel ebben is volt utalás a Gutenberg Projectre, amelynek listáján a Black beauty is szerepelt.) Hasonló szükségtelen többszörözésekhez vezet az eredeti adatbázisokat tükröző Web-helyek átnézése. További gondot okoz az egységesített nevek hiánya (mind a személyek, mind a tárgykörök nevei tekintetében).

A dokumentumok teljes szövegének közvetlen indexelésből adódó problémák egy része áthidalható a már említett TEI rekordfej révén, amelyet a dokumentum létrehozója alkotna meg, s ily módon a szövegbe ágyazottan kereshető lenne. A feladat olyan adatminimumot találni, amelynek alkalmazásával nagy mennyiségű dokumentumot lehet feldolgozni. 1995 márciusában egy Metadata szemináriumot (workshop) tartottak, amely „Dublin Core” néven a következő elemeket határozta meg minimumként: tárgykör, cím, szerző, kiadó, egyéb közreadó (Other Agent), dátum, objektumtípus, forma, azonosító, kapcsolat (Relation), forrás, nyelv és lefedés (Coverage). Az adatminimum elfogadásának több előnye lenne: a szerzőket meta-adatok nyújtására ösztönözné, olyan hálózati publikációs eszközök kifejlesztésére vezetne, amelyek sablont foglalnának magukban a meta-adatok rögzítésére, s ez az adatminimum megkönnyítené a teljes bibliográfiai rekord elkészítését (amennyiben igény van rá). Végül, ha ilyen, minimumra törekvő szabványokat sikerülne elfogadtatni az Internet

résztevőivel, közelebb kerülnének a „lazán összekapcsolt rendszer” megvalósításához, amelyben minden résztvevő tovább dolgozhat a saját eszközeivel, miközben a részrendszereken átívelő struktúrák jönnek létre a közösen használt adatok cseréjére, konverziójára és rétegezésére.

A következő lépések közül a szerző megemlíti a „relációs MARC” megteremtését, amely a Z39.50-es Web-interfészen keresztül hipertext környezetben is működne, s ily módon egyszerű rákattintással más forrásokra vagy az eredeti dokumentumra is át lehetne kapcsolni.

A szervezés harmadik szintje a Metakatalógus lehet, amely képes értelmezni és konvertálni a MARC rekordokat, a TEI rekordfejeket, az SGML, HTML kódokat, beépített nyelvi fordítóprogrammal rendelkezne (az idegen nyelvű kereséshez, tallózáshoz), egységes nevek listáit, valamint többretegű tezaszót tartalmazna (az eltérő szakmai csoportok igényeihez igazodva), s egészében véve zökkenőmentesen tárná fel a releváns információ egészét, nyelvtől és formátumtól függetlenül.

---

A KANADAI KÖNYVTÁRAK DIGITALIZÁLT ÁLLOMÁNYÁT mérte fel 1997 elején a nemzeti könyvtár. A megkérdezett 112 intézmény közül 50 válaszolt, közülük 33 foglalkozik saját dokumentumainak digitalizálásával, 23 gyűjt máshonnan szerzett digitalizált anyagot, 40 pedig a Web útján teszi hozzáférhetővé olvasói számára más könyvtárak ilyen dokumentumait. – Jelentős digitalizált állománnyal még egyikük sem rendelkezik. Főként teljes könyveket, fénykép- és diaanyagot, folyóiratcikket digitalizálnak, téma szerint a helytörténeti anyag és a könyvtár saját állományára vonatkozó tájékoztatói eszköz kap elsősorban ilyen formát. 9 könyvtár csak ideiglenes használatra, 21 viszont archiválásra szánja ezt az anyagot. 28 könyvtárnak van már konkrét terve a munka folytatására. – Külső forrásból leginkább statisztikai és pénzügyi adattárak, referenszművek digitalizált változatát veszik át. 5 könyvtár teljes egészében, 13 pedig részben feltünteteti online katalógusában az ilyen anyagot.  
(*National Library News*, 1997. június).