Aggregált gyűjtemények anatómiája: a Google Print Library Project példája

Bevezetés

A *Google* 2004. december 14-i nevezetes kezdeményezése öt megagyűjtemény nyomtatott könyvállományának digitalizálásáról, és a Google-keresésbe való bevonásáról egyes vélemények szerint üdvös, hiszen kitágítja a könyvtári gyűjtemények láthatóságát, mások ellenben arra figyelmeztetnek, hogy a digitalizálási projekt révén a Google egyfajta portálszerephez jut, és a cég ellenőrzése alá kerülhetnek a könyvtári gyűjtemények. A *Google Print Library Project* (*GPLP*) által generált vita ráirányította a figyelmet arra, hogy mi lesz végső soron a nyomtatott könyvállományok sorsa. A tanulmány szerzői, az *OCLC* munkatársai úgy érvelnek, hogy a könyvbeszerzési források csökkenése, a használói igények eltávolodása a nyomtatott könyvektől, illetve a gyűjtőköri stratégiák könyvtártípusonként eltérő jellege miatt szükség lenne egy egyesített, intézményközi könyvgyűjteményre, a nyomtatott könyvek depójára. A GPLP-hez hasonló tömeges digitalizálási programok némiképp rávilágítanak a könyvgyűjtemények jövőjével kapcsolatos kérdésekre, de a tanulságok levonása ebben a szakaszban még korai volna.

A tanulmány górcső alá veszi a Google öt partnerének (a továbbiakban: *Google 5*) könyvállományát, és összeveti az OCLC egyesített világkatalógusának, a *WorldCat*nek az állományával a következő szempontok alapján:

- *Lefedettség*: A könyvgyűjtemények teljes rendszerének mekkora hányadát fogja a GPLP lefedni? Milyen fokú az átfedés az öt könyvtár állománya között?
- *Nyelv:* Milyen a nyelvek megoszlása a GPLP-ben részt vevő könyvtárak könyvállományában?
- *Szerzői jog*: A GPLP által érintett nyomtatott könyvek mekkora hányada nem esik szerzői jogi védelem alá?
- *Művek*: Hány különböző mű található a GPLP-könyvtárak állományaiban?
- *Konvergencia*: Milyen lenne a lefedettség másik öt könyvtár esetében? Milyen hatása lesz, ha további könyvtárak állományaival gyarapítjuk a *Google 5* könyvtárak egyesített gyűjteményét?

A tanulmány célja néhány alapvető kérdés megfogalmazása, és ezáltal egyfajta tapasztalati kontextus megteremtése a további vitához. Másodlagos cél egy olyan általános kérdéshalmaz megfogalmazása, amely hasznosnak bizonyulhat más tömeges digitalizálási kezdeményezések szempontjából. Itt jegyezzük meg, hogy a tanulmány megjelenése óta a GPLP nevet 2005 novemberében *Book Search*re változtatták, s 2006 augusztusában a *Google ötökhöz* egy hatodik is csatlakozott, az *University of California* 20 millió kötetes könyvtári hálózata; 2006 szeptemberében pedig nyilvános szolgáltatásként megindult a Google Book Search béta-verziója (– A ref.).

A források

Az OCLC WorldCat bibliográfiai adatbázisa stratégiai fontosságú forrás, az egyedüli olyan adatbázis, amely minden részletre kiterjedő információkat tartalmaz az egyes könyvtári gyűjteményekről. Az ismertetendő elemzés a WorldCat 2005. januári állományán alapul, amely kb. 55 millió rekordot tartalmazott. Felhasználták a WorldCat 2005. januári állományfájlját is, amely kb. 1 milliárd példányrekordot foglalt magában.

Az összesített könyvgyűjtemény

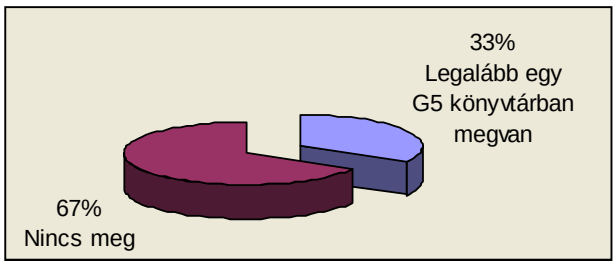
A Google 2004. decemberi bejelentése szerint a cég a *Google 5* könyvtárakkal együttműködésben elkezdi az említett könyvtárak állományába tartozó *könyvek* digitális szkennelését. A GPLP-vel kapcsolatos vizsgálódás tárgya tehát a *könyv*, illetve a *nyomtatott könyv,* s a jelen elemzés a *Google 5* gyűjteményeiben található nyomtatott könyvekre

terjed ki. 2005 januárjában, kb. egy hónappal a Google bejelentése után a WorldCat adatbázisában 32 millió nyomtatott könyv adatai szerepeltek, ez durván a teljes adatbázis 60 százaléka. Látható, hogy a nyomtatott könyvek a könyvtári gyűjtemények jelentős hányadát teszik ki, legalábbis a WorldCatben ez tükröződik. A 32 millió könyvet tartalmazó WorldCatet *Schonfeld* és *Lavoie* az OCLC rendszerére utalva „az egész rendszerre kiterjedő könyvgyűjteményként" aposztrofálja, amely a világ legnagyobb közös katalógusaként az összes – kb. 20 ezer – részt vevő könyvtár egyesített (aggregált) könyvállományát tárja fel.

Lefedettség

A GPLP kapcsán felmerül a kérdés, hogy az összesített könyvgyűjteményt milyen részben fedi le a projekt. A témával foglalkozó összes vita spekulatív jellegű ezen a ponton, mivel egyelőre nem tudni, hogy az egyes könyvtárak állományainak mekkora hányada lesz digitalizálva. Felvázolhatunk ugyanakkor néhány szempontot a GPLP által nyújtott lehetséges legnagyobb lefedettségből kiindulva, azt feltételezve, hogy a részt vevő könyvtárak teljes könyvgyűjteményét digitalizálják, és ezt összehasonlíthatjuk az összesített könyvgyűjteménnyel, amelyet adott esetben a WorldCatben található 32 millió katalogizált nyomtatott könyv reprezentál.

2005 januárjában a WorldCatben a *Google 5* könyvtárak 18 milliós könyvállománnyal képviseltették magukat, vagyis egy könyvtárra 3,6 milliós állomány jutott. Ebből következik, hogy a GPLP digitalizálása a WorldCatben található katalogizált nyomtatott könyvek 57 százalékát fedi le, ha azzal a nem reális feltételezéssel élünk, hogy nincs átfedés az öt részt vevő könyvtár gyűjteménye között. A valóságban természetesen van átfedés, amelyet figyelembe véve az *1. ábrán* szereplő adatokhoz jutunk.

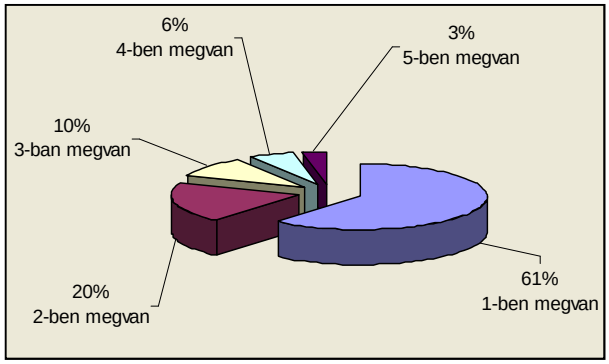


1. ábra **Az összesített könyvgyűjtemény lefedettsége a *Google 5* könyvtárak gyűjteményeiben**

Az összesített gyűjtemény GPLP általi lefedettségének aránya körülbelül egyharmad (33%), vagyis 10,5 millió a 32 millióból. Az összesített gyűjtemény kb. kétharmada, mintegy 21,6 millió könyv tehát egy könyvtár állományában sem található meg az ötből.

A *2. ábra* az állományok közötti átfedést mutatja a közös gyűjteményben található 10,5 millió könyvre vonatkozóan. Láthatjuk, hogy a könyvek mekkora hányada található meg mindössze egy, illetve kettő, három, négy vagy öt *Google 5* könyvtárban.

A GPLP közös gyűjteményében őrzött 10,5 millió könyvből 6,3 millió (61%) található meg csupán az egyik könyvtár állományában az öt közül; 2,1 millió (20%) két, 1,1 millió (10%) három, 0,6 millió (6%) négy, 0,4 millió (3%) pedig öt könyvtár állományában is szerepel. Mindebből következik, hogy ha az

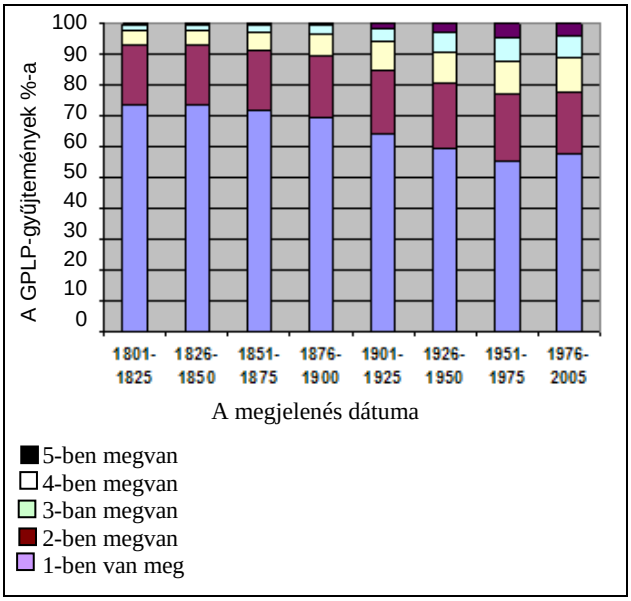


2. ábra **Átfedés a *Google 5* könyvtárak állományai között**

összes gyűjteményt teljes egészében digitalizálják, körülbelül 10-ből 4 könyvet legalább egyszer − fölöslegesen − újradigitalizálnak, vagyis minimum 40%-os redundanciával kell számolnunk. A redundancia mértékét − az OCLC *Functional Requirements for Bibliographic Records (FRBR)* elnevezésű modelljének meghatározásával élve − mutató eredmények az ún. „*nyomtatott könyv megjelenési formákra*" vonatkoznak, ahol a megjelenési forma „egy mű kifejezési formájának fizikai megtestesülése". Ezek szerint pl. *Dickens „Két város regénye"* c. művének két különböző kiadását két különböző könyvnek kell tekinteni. Ha megjelenési formák helyett címekben vagy művekben gondolkozunk, a redundancia foka még magasabb lehet.

Más szemszögből az átfedettség ilyen szintje alacsonynak is mondható. A redundancia foka a kombinált könyvgyűjtemények számának függvénye: minél több a gyűjtemény, annál nagyobb a

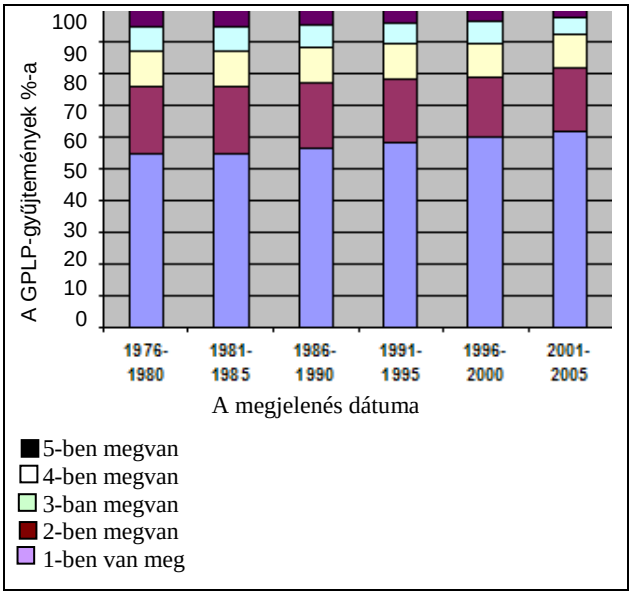
redundancia. Ha azonban az átfedést csak kétoldalú összehasonlítás szintjén vizsgáljuk, egészen más képet kapunk. A legmagasabb redundancia két GPLP-könyvgyűjtemény függvényében 21%, a legalacsonyabb 14%; az átlag 18% körül van. Ebből következik, hogy ha bármelyik két *Google 5* könyvtárat vesszük − illetve ha a *Google 5* könyvtárak eredményeit rávetítjük bármely két nagy kutatókönyvtárra −, 10-ből 8 könyv az egyesített gyűjteményből unikális lesz abban az értelemben, hogy csupán egyetlen könyvtárban található belőle példány. Természetesen az efféle értelmezés kissé elnagyolt, és óvatosan kell eljárnunk, ha bárminemű határozott következtetést kívánunk levonni belőle. Ugyanakkor úgy tűnik, hogy markánsan hitelteleníti azt az álláspontot, amely szerint a kutatókönyvtárak gyűjteményei kevésbé egyediek. A redundanciahányados megfelelő értelmezését ugyancsak megnehezíti, hogy az állományok közötti átfedés mértéke gyakran a könyvek korának függvénye. A *3. ábra* a *Google 5* állományai közötti átfedést illusztrálja az 1800-as évtől 2005-ig, 8 periódusban.



3. ábra **A *Google 5* könyvtárak állományai közötti átfedés a könyvek megjelenési évei szerint (1801−2005)**

Az ábra jól mutatja, hogy a GPLP közös gyűjteményében megtalálható könyvek közül azoknak az aránya, amelyek csupán egyetlen állományban szerepelnek, a könyvek korának csökkenésével az 1801 és 1825 között megjelent könyvek által képviselt 74%-ról az 1951 és 1975 között megjelentek által képviselt 55%-ra csökkent. Vagyis az átfedés az új könyveknél nagyobb, mint a régiekénél. Érdekes ugyanakkor, hogy a legutolsó periódusban (1976−2005) az unikális könyvek aránya 58%-ra növekedett. Ez a jelenség mélyrehatóbb vizsgálódást igényel (*4. ábra*).

Az egyetlen könyvtárban meglévő könyvek aránya az 1976 és 1980, valamint az 1981 és 1985 közötti periódusban volt a legalacsonyabb: 55%. A további időszakokban egyenletesen emelkedett: 1986−1990 között 56%, 1991−1995 között 58%, 1996−2000 között 60%, 2001−2005 között 62%. E tendencia egyik magyarázata a késedelmes szerzeményezés lehet, habár úgy tűnik, hogy az csak az 1995 és 2005 közé eső periódusra vonatkozik. A másik lehetséges magyarázat a gyűjteményezési döntések növekvő eltérése-divergenciája a *Google 5* könyvtárakban. A szerzők mindebből egyelőre csak azt az óvatos következtetést engedik meg maguknak, hogy a GPLP közös gyűjteményében található könyvek kora és az állományok közötti átfedés, s így a digitalizálási redundancia között fennálló fordított arányosság az utóbbi húsz esztendőben érvényét látszik veszíteni.



4. ábra **A *Google 5* könyvtárak állományai közötti átfedés a könyvek megjelenési évei szerint (1976−2005)**

Nyelv

A GPLP bejelentése után többen hangot adtak aggodalmuknak, hogy a digitalizálással létrejövő globális információforrásban az angol nyelvű könyvek fognak dominálni. E félelmek olyan komolynak

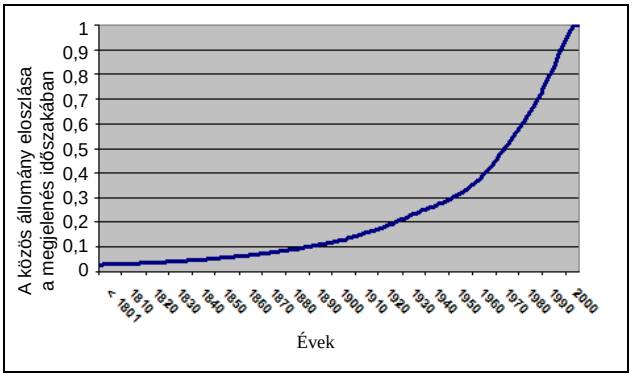
mutatkoztak, hogy 19 európai nemzeti könyvtár egyezményt írt alá egy olyan program létesítése érdekében, amelynek kizárólagos célja a „kontinensünk örökségét képező művek" digitalizálása. Érdemes tehát megvizsgálni a nyelvek eloszlását a *Google 5* könyvtárak közös gyűjteményében szereplő könyvekre, valamint az összesített gyűjteményre vonatkozóan. Megjegyzendő, hogy a WorldCat mint az észak-amerikai könyvtárak közös katalógusa elsősorban észak-amerikai (vagyis angolcentrikus) gyűjteményeket tükröz, s a világ gyűjteményeinek összességéhez képest az angol nyelvű anyag arányaiban nyilván nagyobb.

430 nyelvet azonosítottak a *Google 5* könyvtárak közös gyűjteményében, amelynek valamivel kevesebb mint a felét teszik ki az angol nyelvű könyvek. A gyűjteménynek kb. az egynegyede német, francia és spanyol nyelvű, a maradék a többi nyelvek között oszlik meg. Az összesített gyűjtemény vizsgálata ehhez a megoszláshoz hasonló eredményt mutat. Az a tény, hogy a közös gyűjtemény négy amerikai és egy brit könyvtár állományából tevődik össze, első pillantásra ellentmondásban van azzal, hogy az angol és nem angol nyelvű könyvek aránya közel egyenlő (50-50%). A magyarázat az állományok összeadásában (idegen szakkifejezéssel: aggregálásában) rejlik. Az angol nyelvű nyomtatott könyvek aránya egy angol ajkú országban átlagosan magasabb, mint a nem angoloké: nagyjából 70−75%. Amikor azonban a gyűjteményeket összeadjuk, az állományban nagyobb átfedést tapasztalunk az angol nyelvű könyvek, mint a nem angol nyelvűek között. Ha tehát a duplikátumokat töröljük, nagyobb arányban törlünk angol nyelvűeket, a nem angol nyelvű könyvek aránya viszont nő a közös gyűjteményben az egyes gyűjteményekhez képest. Ez a jelenség további gyűjtemények hozzáadásakor még hangsúlyosabbá válik.

Némiképp megerősíti ezt a magyarázatot, ha megvizsgáljuk a *Google 5* könyvtárak közös gyűjteményében az állományok közötti átfedést az angol és a nem angol nyelvű nyomtatott könyvekre kivetítve. A nem angol nyelvű könyvek 63%-a található egyetlen könyvtárban, míg az angol nyelvűek 57%-a tartozik e kategóriába. A nem angol nyelvű könyvek csupán 6%-a található meg legalább négy könyvtárban, míg ez az angol nyelvű könyvek 13%-ára igaz. Összefoglalva: az angol könyveknél nagyobb állománybeli átfedést tapasztalunk, mint a nem angolokénál, amely tényállás növeli az utóbbiak arányát a közös katalógusban, ha eltávolítottuk a duplikátumokat. Felmerül ennek kapcsán a kérdés, hogy az európai digitalizálási egyezményt jegyző könyvtárak félelmei mennyire megalapozottak. A *Google 5* gyűjtemény valóban angolcentrikus, lévén a gyűjtemény közel fele angol nyelvű, ám sokak véleménye szerint ez az arány rendkívül alacsony. (Ezt tovább árnyalja, hogy az angol nyelvű könyvek egy része fordítás más nyelvből.) Végül megállapíthatjuk, hogy a több mint 400 nyelv jelenléte a könyvállományban azt sugallja, hogy a GPLP által létrehozott információforrás a várhatónál jóval nagyobb mértékben tükröz kulturális sokszínűséget.

Szerzői jog

A GPLP-hez hasonló tömeges digitalizálási programoknál elkerülhetetlenül felmerülnek különböző, a szellemi tulajdonnal kapcsolatos problémák. 2005. augusztus 11-én a Google bejelentette, hogy ideiglenesen felfüggeszti a szerzői jogok által érintett könyvek digitalizálását, lehetőséget adva a kiadóknak: döntsék el ők, hogy mely könyveket szeretnének, illetve nem szeretnének bevonni a programba. Ez az intézkedés, valamint a szerzői jogok megsértéséről és a méltányos használatról folytatott heves vita azt sugallja, hogy érdemes elemezni a *Google 5* könyvtárak közös gyűjteményében lévő könyvek megjelenési adatait. Az *5. ábra* mutatja a *Google 5* könyvtárak közös állományának (10,5 millió könyv) kumulatív eloszlását a megjelenési időpont függvényében.



5. ábra **A *Google 5* könyvtárak közös gyűjteményében található könyvek kumulatív eloszlása**

A *Google 5* könyvtárak közös gyűjteményében található nyomtatott könyvek mintegy fele 1974 után jelent meg, csaknem háromnegyed részük a második világháború után. Ha az 1923-as esztendőt tekintjük határértéknek a szerzői jog alá eső

könyveknél − vagyis abból indulunk ki, hogy az USA copyrighttörvénye alapján az 1923 előtt megjelent könyvek nem esnek szerzői jogi védelem alá −, akkor a *Google 5* könyvtárak közös gyűjteményében található anyag 80%-áról állapíthatjuk meg, hogy *a szerzői jogvédelem hatálya alá esik*. Az összesített könyvgyűjteményben található 32 millió könyv kumulatív eloszlása a megjelenések időpontjának függvényében közel azonos a *Google 5* gyűjteményében található könyvek ugyanilyen szempontból vizsgált eloszlásával, azzal a parányi különbséggel, hogy a *Google 5* könyvtáraknál az eloszlás a 20. század korai éveitől kezdve előrefelé haladva enyhén meredekebben emelkedik.

Az összesített könyvgyűjteményben mintegy 5,4 millió olyan könyv található, amely nem esik szerzői jogi oltalom alá. Körülbelül egyharmaduk található meg a GPLP-ben részt vevő öt könyvtár közül legalább az egyikben. Érdekes módon a *Google 5* könyvtárak és az összesített gyűjtemény a szerzői jog által érintett könyveknél ugyanakkora arányszámot mutat, jóllehet az állománybeli átfedés a szerzői jogon kívül eső könyveknél a *Google 5* könyvtárakban szignifikánsan kisebb: a köztulajdonú (public domain) könyvek több mint 70%-a csupán egyetlen könyvtár állományában szerepel, míg az összesített könyvgyűjteményben ez az arány 60%.

Némi eltérés tapasztalható az öt könyvtár között a szerzői jogi védelem alá nem eső könyveknek a teljes állományhoz viszonyított arányát illetően. Három könyvtárban ez az arány 10% körül mozog, a másik két könyvtárban viszont ennek közel kétszerese: 18%. Egyrészt jelentős különbségek lehetnek a nagy kutatói könyvtárak könyvgyűjteményei között a szerző jogi oltalom alá nem eső könyvek számát tekintve, másrészt a szellemi tulajdonra vonatkozó jogi szabályozások könyvtártól függően eltérően hatnak a tömeges digitalizálási programokra.

A szerzői jogi védelem alá nem eső könyvek arányának kiszámolásakor az 1923-as dátumnál húzták meg a határt; ez az arány tehát *alsó határként* fogható fel a valóságos értékekhez viszonyítva. Ami az 1923 és az 1963 közötti éveket illeti, a szerzői jogról szóló törvény szerint az ekkor publikált könyvekre 28 évig volt érvényes a szerzői jogi védelem, amely további 47 évvel volt meghosszabbítható (amely 47 év 67-re emelkedett a jelenlegi törvény értelmében). Ha a szerzői jogot nem újították meg, a könyvek közkinccsé válnak. Ha azzal a − természetesen hamis − feltételezéssel élünk, hogy az ebben az időszakban megjelent könyvek egyikének sem hosszabbították meg a szerzői jogát, a szerzői védelem alá nem eső könyvek arányának *felső határát* kapjuk meg, az 1963-as dátumot véve határértéknek.

Ha újra megvizsgáljuk az 5. ábrát, ezúttal feltételezve, hogy az 1963 előtt megjelent könyvek egyike sem esik szerzői jogi védelem alá, az előbbitől eltérő kép tárul elénk, ami a szellemi tulajdonnak a javasolt digitalizálásra kifejtett hatását illeti. Az 1963-as esztendőt használva határértékként, a *Google 5* könyvtárak közös gyűjteményének kb. 63%-a esik szerzői jogi oltalom alá; ez az előzőhöz (80%) képest lényegesen kisebb arány. Az összesített könyvgyűjtemény esetén ez a szám 66%, szemben a 80%-kal az 1923-as határértéknél.

Az 1963-as évet véve határértéknek, az összesített könyvgyűjteményben 10,5 millió olyan könyvvel számolhatunk, amely nem esik szerzői jogi védelem alá. Ezeknek a könyveknek mintegy 36%-a megtalálható a *Google 5* könyvtárak közül legalább egyben. Ez az arány csak kicsivel magasabb, mint amikor az 1923 előtti könyveket ítéltük az előbbi halmazba tartozónak. Az állományok közötti átfedést is figyelembe véve ennél nagyobb eltérést regisztrálhatunk: az 1963 előtti könyvek 65%-a található meg csupán egyetlen könyvtárban, szemben az 1923 előttieknél kalkulált 70%-kal. A szerzői jogi védelem alá nem eső könyvek könyvtárankénti aránya az 1963-as határértéknél jóval nagyobb, mint az 1923-asnál, jóllehet az eltérési minta hasonló. Három könyvtárnál ez az arány (a teljes állományhoz mérten) kb. 28%, kettőnél ennél jóval magasabb: 37 és 40%.

Az összesített könyvgyűjteményben található szerzői jogi védelem alá eső könyvek aránya tehát az 1923-as és 1963-as határértékkel valahol 66 és 82% között van; a valós arányt akkor állapíthatjuk meg, ha megtudjuk, hogy az 1923 és 1963 között megjelent könyvek közül hánynak újították meg a szerzői jogát. Röviden: a *Google 5* könyvtárak közös gyűjteményének *legalább* egyharmadát védi a szerzői jog, jóllehet a szerzői jogi korlátozások a GPLP könyvtárakat különböző mértékben érintik; ha az 1923-as esztendő a határérték, az állomány érintettsége 82 és 90% között van, az 1963-as határértékkel számolva 60 és 72% között mozog.

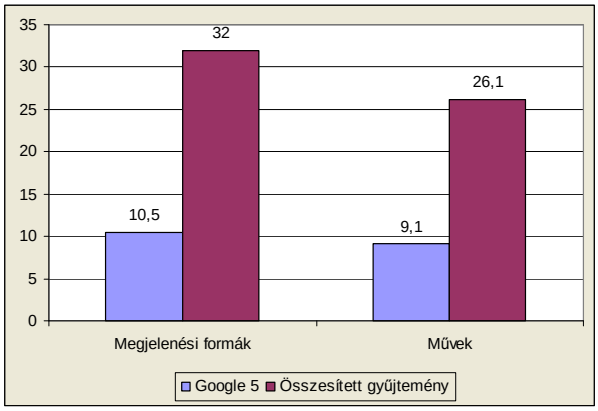
Művek

Az OCLC FRBR-modelljének meghatározása szerint a *mű* „önálló szellemi vagy művészi alkotás"; Shakespeare *Macbeth*je tehát műnek tekinthető. Valamely mű *kifejezési formája* „a mű szellemi vagy művészi megvalósítása alfanumerikus, zenei vagy koreográfiai jelölési rendszerben, zenei, képi, tárgyi, mozdulati stb., vagy mindezen formák bármely kombinációjaként". A Macbethnek egy angol nyelvű szövege a Macbeth c. mű kifejezési formája. A *megjelenési forma* (manifesztáció): „a mű kifejezési formájának valamilyen fizikai megtestesülése". A *Folger Shakespeare Könyvtár* gondozásában, a *Washington Square Press* kiadásában, 2004-ben puha kötésű könyv formájában megjelent Macbeth a Macbeth c. mű egyedi megjelenési formája.

A WorldCat-rekordok általában megjelenési formákat írnak le, és a már bemutatott eredmények is ezekre vonatkoznak. Ugyanakkor könnyen elképzelhető, hogy vannak olyan körülmények, amelyeknél a használói igényeknek jobban megfelel, ha egy magasabb szintű bibliográfiai entitás („kifejezési forma", „mű") a digitalizálás tárgya. A Google kezdeményezése megjelenésiforma-példányok digitalizálására irányul.

Az OCLC kidolgozott egy algoritmust MARC21 alapú bibliográfiai adatbázisok FRBR-*műhalmazzá* való konvertálására. Műhalmaznak a WorldCat-rekordok − megjelenési formák − olyan csoportját tekintjük, amely egy és ugyanazon műnek felel meg. Az összesített könyvgyűjtemény 32 millió manifesztációja 26,1 millió különálló műre vonatkozik. Minden egyes műre 1,2 nyomtatottkönyv-manifesztáció jut, vagyis egy műre egy nyomtatott könyv. A *6. ábra* a *Google 5* könyvtárak megjelenési formákra és művekre vonatkozó lefedettségi adatait tartalmazza.

A 26,1 millió különálló nyomtatott mű közül 9,1 millió, 35% található meg legalább egy GPLP-könyvtárban, ami jelzi, hogy a művek lefedettsége a csak kicsivel nagyobb, mint a megjelenési formáké. A művek 56%-a található meg egyetlen *Google 5* könyvtárban, ez az adat a megjelenési formák tekintetében 60%. Ebben nincs semmi meglepő, hiszen a megjelenési formáknak művekként való csoportosítása csökkenti a gyűjtemények egyediségét. E csökkenés nem túl jelentős, mivel a legtöbb műnek csupán egy, legfeljebb néhány megjelenési formája létezik. Ami az állományeloszlást illeti: a művek kb. 12%-a található meg legalább négy *Google 5* könyvtárban, szemben a megjelenési formák 9%-ával.



6. ábra **A *Google 5* könyvtárak megjelenési formákra és művekre vonatkozó lefedettségi adatai (millió)**

A művek 44%-a található meg két vagy több *Google 5* könyvtárban, amiből következik, hogy a *Google 5* könyvtárak teljes gyűjteményének digitalizálása esetén 10-ből több mint 4 könyv digitalizálása fölösleges volna, ha feltételezzük, hogy a *művek (címek*), s nem a megjelenési formák digitalizálása a projekt célja. Látszólag hasonló redundanciafokkal kell számolnunk a megjelenési formák digitalizálása esetén, mivel − mint említettük − a legtöbb műnek csak egy-két megjelenési formája van. Az eredmények azonban elfedik azt a tényt, hogy valószínűleg létezik a sok állományban szereplő, számos megjelenési formában meglévő műveknek egy „maghalmaza", amelynek következtében a redundanciahányados rendkívül magas lesz. Ezért vezethet jelentős költségmegtakarításhoz, ha a megjelenési formák helyett a művek vagy kifejezési formáik digitalizálására összpontosítunk.

Konvergencia

A GPLP-t pozitív kezdeményezésként értékelők a projekt egyik érdemének azt tekintik, hogy az első lépést jelentheti a világ összes könyvtárában fellelhető könyvgyűjtemények digitalizálása és internetes (online) hozzáférhetővé tétele felé. Jóllehet e cél elérése nem tűnik túl egyszerűnek. *Shonfeld* és *Lavoie* nemrég megjelent cikkükben azt írják, hogy a WorldCatben összesített könyvgyűjtemény rengeteg intézmény között oszlik el. *A nyomtatott könyveknek közel 40%-a csak egyetlen intézményben található meg!* A könyveknek csak har-

mada található meg több mint öt állományban, s kb. fele kettőben vagy egyben. Vagyis az összesített könyvgyűjtemény valóban sok intézmény között oszlik meg, és sok könyv számít ritkának abban az értelemben, hogy kevés intézmény állományában lelhető fel.

A GPLP – mint láttuk – az összesített könyvgyűjtemény kb. egyharmadát fedi le. Ilyen fokú lefedettség elérése mindössze 5 könyvtár állományának egyesítésével jelentős eredménynek számít, ugyanakkor felvet két kérdést: (*1)* Milyen eredményre jutnánk, ha másik öt könyvtár venne részt a programban? (*2)* A lefedettség milyen mértékű növelését érnénk el további könyvtáraknak az eredeti öthöz való hozzáadásával? Ezek megválaszolására találomra kiválasztottunk további öt könyvtárat: egy kis amerikai bölcsészettudományi főiskola, egy nagy kanadai egyetem, egy nagy amerikai állami egyetem, egy nagy amerikai magánegyetem könyvtárát, és egy nagy amerikai városi könyvtárat. Az öt új könyvtár egyesített állományában 5,9 millió nyomtatott könyv van, vagyis az egész rendszerre kiterjedő nyomtatott könyvgyűjteményben található 32 millió 18%-a. Ez jóval kevesebb, mint a *Google 5* könyvtárak egyesített gyűjteményében szereplő 10,5 millió könyv, de ha az eredményeket kiigazítjuk az ezen állományok és a *Google 5* könyvtárak állományai közötti méretbeli egyenlőtlenséggel, más képet kapunk. Az egyetlen állományban található könyvek aránya a teljes állomány 74%-a az új könyvtárak, és 58%-a a *Google 5* könyvtárak esetében. Ez azt jelenti, hogy a közös gyűjtemény esetén kisebb redundanciával kell számolnunk: a *Google 5*-nél 10-ből négy könyv digitalizálása volna fölösleges, míg az új egyesített gyűjteménynél mindössze kettőé vagy háromé.

A redundancia alacsonyabb foka következik az állományeloszlás vizsgálatából is. Az 5,9 millió könyv közel háromnegyede található meg csupán egyetlen könyvtárban, ugyanez az arány a *Google 5* könyvtárak esetén 60%. A nyomtatott könyvek 9%-a található meg legalább négy *Google 5* könyvtárban, az új gyűjtemény esetén ez az arány mindössze 1%. Ha az öt új könyvtár gyűjteményeit egyenként összevetjük a *Google 5* egyesített gyűjteményével, megvizsgálhatjuk, hogy milyen hatással van a lefedettségre, ha különböző profilú könyvtárak állományaival bővítjük a közös gyűjteményt. A nagy amerikai magánkönyvtár nagyszámú, mintegy 1 millió egyedileg őrzött példánya 10%-kal növeli meg a *Google 5* közös állományát. A kis amerikai bölcsészettudományi főiskola 71 ezer egyedileg őrzött könyvével kevesebb mint 1%-os állománynövekedést, a nagy amerikai állami egyetem könyvtára közel félmillió könyvével 5%-ost okozna, a nagy városi könyvtár több mint 231 ezer könyvével 2%-ost, a nagy kanadai egyetemi könyvtár kb. 104 ezer könyvével 1%-ost.

Ezek az eredmények részben a gyűjteményméretek egyenlőtlenségének következményei: a nagy amerikai magánegyetemnek van a legnagyobb állománya, a második legkisebb pedig a bölcsészettudományi főiskolának. A gyűjteményméretre vonatkozó adatokat pontosíthatjuk, ha megvizsgáljuk, hogy azoknak az egy könyvtár által őrzött könyveknek a száma, amelyeket a *Google 5* könyvtárak közös állományához hozzáadunk, hogyan aránylik az egyes intézmények teljes állományához. Ebből a szempontból a nagy amerikai városi könyvtár éri el az állomány egyedisége tekintetében a legnagyobb százalékot: 39% a *Google 5* állományában nem található egyedi könyvek aránya a teljes állományhoz képest. A nagy amerikai magánegyetem 25%-kal a második, ezt követi a kanadai egyetemi könyvtár (23%), a nagy amerikai állami könyvtár (21%), és a kis amerikai bölcsészettudományi főiskola (12%).

A szerzők végül összehasonlították a *Google 5* könyvtárak, illetve az újonnan kiválasztott öt könyvtár egyesített gyűjteményeit. Ezekben együttesen 12,3 millió könyv található, vagyis a növekedés 1,8 millió könyv, kb. 17% a *Google 5* könyvtárak közös gyűjteményéhez mérten. Ebből következik, hogy az összesített könyvgyűjtemény digitalizálása sok-sok különböző típusú könyvtár közreműködését igényli: ha ugyanis egy különböző könyvtárakból származó 8 milliós állományt hozzáadunk a *Google 5* könyvtárak egyesített gyűjteményéhez, az így létrejött közös állománynak mindösszesen 8%-a olyan könyv, amely a *Google 5* könyvtárak egyikének állományában sincs meg. Valószínű, hogy ha további öt könyvtárat adunk ehhez a gyűjteményhez, a növekedési arány még kisebb lesz.

Következtetés

Arra, hogy miben rejlik a Google Print Library Project jelentősége, csak a későbbiekben derül fény. Az ismertetett tanulmány néhány olyan területre tér ki, ahol valamilyen hatás várható: lefedettség, nyelv, szerzői jogok, művek, konvergencia. A cikk másik erénye, hogy egyfajta „tapasztalati kontextussal" szolgál a vonatkozó kérdések továbbgondolásához.

A GPLP-hez hasonló célokat megfogalmazó projektek szaporodásával egyre hasznosabbá válik egy több intézményre kiterjedő tömeges digitalizálási programokra vonatkozó általános kérdéshalmaz megfogalmazása:

- Milyen jellegzetességei vannak a digitalizálás tárgyát képező anyagok „populációjának"?
- A „populáció" mekkora hányadát fedi le potenciálisan a digitalizálás?
- Milyen redundanciafokkal kell számolni a digitalizálásnál?
- Mely bibliográfiai egység (pl. megjelenési forma, kifejezési forma, mű) áll a digitalizálás középpontjában?
- Hány részt vevő intézmény, és a különböző intézménytípusok milyen egyesítése lenne optimális ahhoz, hogy a lehető legkisebb befektetéssel a lehető legnagyobb haszonra tegyünk szert, ami a digitalizálás során kitűzött célok egy részhalmazának elérését illeti?

A digitalizálási programok elterjedésével a legtöbb kezdeményezés valószínűleg a könyvtári közösségekből származik majd, s nem annyira a Google-hoz hasonló külső szervezetekből. A könyvtári kezdeményezésű és alapítású programoknál különösen fontos, hogy egyrészt a digitalizálást oly módon szervezzék meg, hogy az elérhető források hasznosítása maximális közösségi hasznot hozzon, másrészt a digitalizálás olyan stratégiát tükrözzön, amely számol az egész könyvtári világrendszerre kiterjedő következményekkel. A tervbe vett digitalizálási programoknak a legjobb elérhető adatforrásokra támaszkodó gondos elemzése segítheti a döntéshozókat abban, hogy előre lássák a programok hatásait, és úgy alakítsák őket, hogy hozzájáruljanak mindkét említett cél megvalósulásához.

/LAVOIE, Brian−CONNAWAY, Lynn Silipigni−DEMPSEY, Lorcan: Anatomy of aggregate collections: the example of Google Print for libraries. = D-Lib Magazine, 11. köt. 9. sz. 2005. 15 p. http://www.dlib.org/dlib/september05/lavoie/09lavoie.html
Zeitschrift für Bibliothekswesen und Bibliographie, 52. köt. 6. sz. 2005. p. 299−310./

(Dancs Szabolcs)