megtehetik, hogy manuális módszerekkel állítanak elő bizonyos metaadatokat. A University of California kampányszövegeket gyűjtő archívumánál például Dublin Core adatmezőket, Library of Congress tárgyszavazást és saját besorolási állományokat használnak a katalogizáláshoz. A *Digital Archive for Chinese Studies* sinológusokat kért fel a leíró metaadatok elkészítéséhez. A *National Taiwan University Web Archives* fejlesztői háromszintű osztályozási rendszert és speciális katalogizálási szabályokat dolgoztak ki a webes tartalmakhoz. Más rendszereknél a felhasználók is címkézhetik, kommentálhatják és értékelhetik az archivált anyagokat. A Library of Congress MODS rekordokat készít azokból az adatokból, amelyeket az archiválandó oldalakat javaslók szolgáltatnak, majd ezeket a rekordokat a katalogizálók még kiegészítik és pontosítják.

Gyakori megoldás, hogy előbb a nagyobb egységeket (pl. a webhelyeket) metaadatolják, majd ha van rá ember, akkor weblapszinten is elvégzik a leírást. Fájlszintű katalogizálásra (pl. az oldalakon található minden egyes kép önálló leírására) ritkán van példa, de bizonyos automatikusan generálható metaadatokat (pl. formátum, méret, módosítási dátum) ezen a szinten is elő lehet állítani. Minél kisebb egységet választunk, annál pontosabb leírások készíthetők, és természetesen annál több metaadatrekord fog keletkezni. A *Harvard University* webarchívumánál csak egyetlen, az online katalógusban is visszakereshető MARC rekordot készítenek a könyvtárosok az egyes részhalmazokról, amelyek rendszerint több webhelyből állnak. A Library of Congress hasonlóképpen, részgyűjteményenként katalogizálja az archivált anyagát, de emellett minden website-hoz saját MODS rekord is készül – utóbbiak azonban csak az archívumon belül kereshetők, az OPAC-ban nem jelennek meg. Az ausztrál PANDORA esetében a leírási szint egyaránt lehet a teljes webhely vagy annak valamilyen kisebb egysége.

Hozzáférés és használat

Hogy az archivált tartalomhoz ki és hogyan férhet hozzá, azt elsősorban az adott országban érvényes jogi szabályozás határozza meg. Új-Zélandon nemcsak a publikus weboldalak archiválását engedi meg a kötelespéldány-törvény, hanem az archívum nyilvános szolgáltatását is. Az Egyesült Államokban a Library of Congress csak a bibliográfiai leírásokat teszi teljes körűen visszakereshetővé, nyilvános hozzáférést csak azokhoz a webhelyekhez tesz lehetővé, amelyek tulajdonosai erre engedélyt adtak. Sok webarchívum zárt vagy csupán helyben használható – ilyen például a francia, a finn, a dán, a norvég, a szlovén, a svájci és az osztrák. Más esetekben csak csökkentett funkcionalitással vagy pedig késleltetéssel engedik a nyilvános hozzáférést. A Harvard University Library WAX rendszerénél például legalább 3 hónap a késleltetés, az *IA Wayback Machine* szolgáltatásánál pedig 6-12 hónap a várakozási idő azért, hogy ne jelentsenek konkurenciát az eredeti, „élő" webhelyeknek.

A keresési lehetőségeket az alkalmazott technológia és a metaadatok részletessége határozza meg. A Library of Congress és a National Library of New Zealand archívuma – a *subject headings* szerinti osztályozásnak köszönhetően – authoritylisták segítségével böngészhető. Ezzel szemben a Wayback Machine csak URL cím alapján tud megtalálni egy oldalt. A *NutchWax* keresőgépet használó rendszerek teljes szövegű keresést is biztosítanak. Vannak érdekes vizualizációs kísérletek is: az Egyesült Királyság archívumához adatbányász módszerekkel címkefelhőket készítettek, illetve egy 3D-ben animált falon lehet megnézni az egyes weblapok alakulását az időben. Japán kutatók pedig diavetítés és grafikon segítségével kísérelték meg bemutatni azt, hogy egy URL cím mögött hogyan változik a tartalom.

/NIU, Jinfang: An Overview of Web Archiving. = D-Lib Magazine, 18. köt. 3–4. sz. 2012./

(Drótos László)

A webarchívumok funkcionalitása

A web megőrzésének egyes munkafázisait és a jelenlegi gyakorlatot összefoglaló korábbi cikkét követően a szerző ebben az írásában néhány nyilvános webarchívumot elemez funkcionalitás szempontjából. Ahogy a könyvtárakban és levéltárakban fokozatosan kialakul ennek az állománygyarapítás-fajtának a gyakorlata, remélhetőleg több idő és figyelem jut majd erre a részterületre is, vagyis az

archívumhasználók által igényelt különféle funkciók beépítésére. A cikk mellékletében közzétett, közel negyven funkciót és szolgáltatásfajtát tartalmazó ellenőrző lista segítséget nyújthat a webarchívumokat üzemeltetők számára a jelenlegi rendszerük értékeléséhez és a továbbfejlesztési irányok meghatározásához.

A kutatás ismertetése

Az IIPC Access Working Group, vagyis az internet archiválásával foglalkozó nemzetközi konzorciumnak a hozzáférés kérdésére specializálódott munkacsoportja 2006-ban olyan hipotetikus esetleírásokat fogalmazott meg, amelyek a webarchívumok tipikus felhasználási formáit illusztrálják. [1] Minden ilyen eset többféle funkciót is feltételez: a legegyszerűbb „URL-re való keresés"-től, a legkomplexebb „adatbányászat"-ig. Egy évvel később *Ras* és *Bussel* a holland nemzeti webarchívum felhasználóinak lehetséges típusait és ezek igényeit tanulmányozta [2] és fogalmazott meg különböző szempontokat a kezelő- és keresőfelülettel kapcsolatban. 2010-ben pedig *Costa* és *Silva* tartott egy előadást [3] a portugál webarchívum használóinak keresési szokásairól és elvárásairól. Úgy találták, hogy az archívum esetében sokkal gyakoribb a konkrét webhelyre vagy weblapra való keresés, mint egy adott témával kapcsolatos információgyűjtés; továbbá, hogy a régebbi mentéseket gyakrabban használják, mint az újabbakat.

E két publikációból, valamint az IIPC esetleírásokból leszűrhető igények alapján a szerző egy listát állított össze a webarchívumoktól elvárható funkciókból (pl. keresési és böngészési módok, letiltási vagy bekerülési lehetőség, adatbányászati, illetve webhely-helyreállítási szolgáltatás) és ezt összevetette néhány publikus archívum jelenlegi funkcionalitásával. Az értékelésre kerülő szolgáltatásokat az IIPC nyilvántartásából (*netpreserve.org*) választotta ki. Ez a jegyzék 24 archívumot tartalmaz, közülük az egyik (*Bibliotheca Alexandrina*) az Internet Archive (IA) tükrözése ugyanazokkal a funkciókkal, így ezt nem volt értelme külön értékelni. További nyolc rendszer esett ki azért, mert vagy zárt, vagy csak helyben használható. A maradékból kilencnek van angol felülete, ezekhez még tízediknek érdemes volt hozzávenni az Archive-It szolgáltatást, mert bár ezt is az IA működteti, mint a *Wayback Machine* nevű rendszert, de jelentős a különbség a két szolgáltatás funkcionalitása között. Az Archive-It, akárcsak a *California Digital Library* által indított *WAS* (*Web Archiving Service*), valójában nem önálló archívum, hanem olyan infrastruktúra, amellyel az előfizetők úgy tudnak webarchívumokat építeni, hogy nem kell foglalkozniuk a technikai kérdésekkel, és nem szükséges saját tárolószervert működtetniük. 2011 áprilisában a WAS-nak 16, az Archive-It szolgáltatásnak pedig 160 előfizetője volt (köztük nemzeti és tudományos könyvtárak, levéltárak és kormányhivatalok).

Eredmények és következtetések

Keresési lehetőségek: A leggyakoribb az URL cím alapú hozzáférés, ezt követi a kulcsszavas kereshetőség. A tízből hat archívumnál pontos URL-t kell megadni, a másik négynél keresőkérdésként írhatunk be URL-eket, de ilyenkor olyan találatok is előjöhetnek, amelyeknél az eredeti webcím részeként vagy magán a weblapon fordul elő a keresett URL. A *Library of Congress Web Archives* és a *New Zealand Web Archive* esetében a kulcsszavas keresés nem a teljes szövegben, hanem a bibliográfiai adatrekordokban történik. A vizsgált archívumok felénél lehet doménre, vagyis webhelyre korlátozni a keresést, a PANDORA legfelső szintű doménnévre (pl. *.gov* vagy *.edu*) is tud szűrni. Dátumra szűkítés opció hat rendszerben van, a lépték nagyon különböző: a kanadai és a brit, valamint az amerikai IA keresőjében napra pontos intervallumot állíthatunk be, az Archive-It esetében ez már csak hónapnyi pontossággal lehetséges, az ausztrál PANDORA és a *Library of Congress* keresőjében pedig csak évekre korlátozhatunk. Tízből hét szolgáltatás kínál médiatípus beállítási lehetőséget: van, ahol csak HTML és PDF az alternatíva, de olyan is akad (UK Web Archive), ahol nyolcféle formátum, illetve médiatípus közül választhatunk. A hagyományos könyvtárakhoz kötődő webarchívumok a könyvtári katalógusból is elérhetők. Az Archive-It rendszernél egyszerre lehet keresni mindegyik publikus előfizetői gyűjteményben. A WAS viszont nem kínál ilyen közös keresőt. A *UK Government Web Archive*-ot integrálták a kormányzati portállal: ha egy olyan URL kérés érkezik a szerverhez, amely már nem létező oldalra mutat, akkor a böngészőprogramot automatikusan az archívumba irányítják. Egyik vizsgált webarchívumnál sincs nyoma annak, hogy lehetne MD5 „ujjlenyomat" alapján teljesen azonos másolatokra keresni, vagy műfaj, illetve frissítési gyakoriság szerint szűrni. A felnőtt tartalom kiszűrésére sem kínálnak megoldást ezek a rendszerek (ilyenre persze nincs is mindenhol szükség). Összetett keresőt nem mindegyik szolgáltatás működtet, és

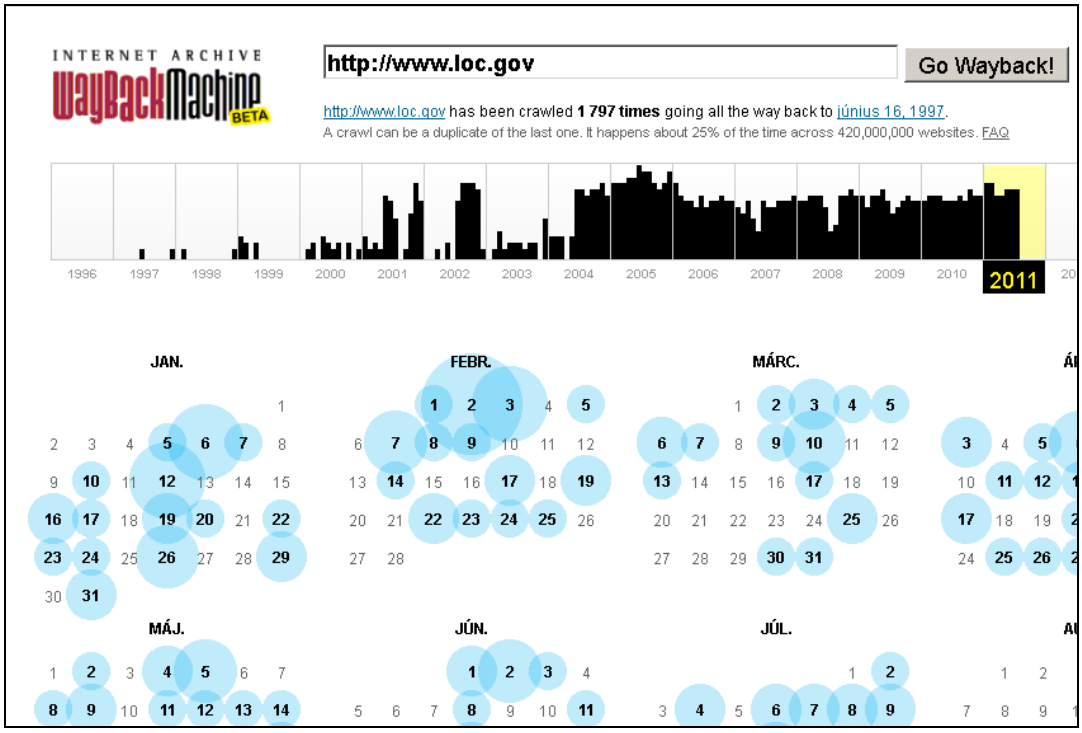
ahol van, ott sem könnyű mindig megtalálni vagy használni.

Találatok: A találati listák mindig tartalmazzák az archivált weblapok URL-jét (csoportosítva az azonos címeket), valamint a mentés időpontját – utóbbit az IA Wayback Machine például naptárszerűen mutatja *(1. ábra)*. Öt archívumnál egy rövid kivonat is látható a weblapok tartalmából, a Library of Congress rendszere és a New Zealand Web Archive viszont a bibliográfiai leírásokat jeleníti meg. Mindegyik archívum képes arra, hogy a felhasználó úgy navigálhasson a mentett anyagban, mint az élő weben.

Stabil azonosítók: Négy rendszer nyújt hosszú távon is állandó azonosítót minden archivált weblaphoz, sőt a PANDORA kisebb egységekhez (pl. egy-egy beágyazott képhez vagy táblázathoz) is képes egyedi azonosítókat generálni. A *Harvard University Library* WAX rendszerénél kész hivatkozásokat tölthetünk le háromféle formában (APA, Chicago és MLA) háromféle szinthez (archív gyűjtemény, webhely, weblap), és a bennük megadott URL egyben stabil azonosítóként is szolgál. A Library of Congress archívumánál a böngésző címsorában megjelenő URL tartalmazza a lementés időpontját és az archivált weboldal eredeti URL-jét is. A legtöbb szolgáltatásnál viszont inkább egy – esetleg kattintással elrejthető – felső sávba írják ki ezeket az információkat, illetve azt, hogy a felhasználó nem élő oldalt, hanem archív példányt lát.

Nyomtatás: Három olyan szolgáltatás volt a vizsgált tízből, ahol ez a felső *banner* sáv nem nyomtatódott ki a weboldallal együtt, kettőnél pedig bizonyos képek is lemaradtak és az eredeti külalak is elveszett.

Hitelesítés: Egyik archívumnál sincs említés az archivált anyagok hitelességének igazolásáról. Inkább az ellenkezőjére vonatkozó állításokkal lehet találkozni (pl. az IA Wayback Machine és a WAX, sőt még a hivatalos kormányzati dokumentumokat gyűjtő *Government of Canada Web Archive* honlapján is), vagyis elhárítják maguktól a felelősséget az archívumban található tartalom pontosságával vagy megbízhatóságával kapcsolatban. A PANDORA ismertetője ugyan azt állítja, hogy különféle módszerekkel törekednek a lementett források hitelességének és integritásának megőrzésére, de arról már nem szól, hogy erre vonatkozóan tanúsítványt lehetne kérni az archívumtól.



1. ábra **Az amerikai Kongresszusi Könyvtár webhelyének mentési naptára és grafikonja az Internet Archive Wayback Machine keresőfelületén**

Böngészés: Nyolc archívumnál van böngészési funkció, ezek közül hétnél részgyűjtemények vagy tematikus, illetve műfaji kategóriák szerint lehet válogatni, a kanadai kormányzati archívum pedig minisztériumok szerint böngészhető. Még a kizárólag automatikus módszerekkel metaadatolt archívumoknál (mint pl. az IA Wayback Machine) is érdemes volna böngészési lehetőséget biztosítani a felhasználóknak, például ország-, illetve legfelső szintű doménnevek, azon belül műfajok (pl. blogok, híroldalak, virtuális világok), illetve médiatípusok (pl. PDF, HTML, videó) szerint. Egy böngészhető hierarchia ugyanis lehetővé teszi olyan források megtalálását, amelyeknek nem tudjuk a pontos URL címét.

Irányelvhez kapcsolódó funkciók: Itt olyan funkciókról van szó, amelyekkel töröltetni lehet webhelyeket/weboldalakat a nyilvános archívumból, vagy kérni lehet ezek archiválását. Ezen a téren meglehetősen eltérő a gyakorlat. Csak a WAS rendszernél említik kifejezetten a letiltás lehetőségét, de másik hatnál is van valamilyen *„takedown policy”*, vagyis kérésre el lehet távolíttatni anyagokat. A Library of Congress eleve blokkolja azokat a site-okat a nyilvános felületen, amelyekre nem kapott előzetesen engedélyt a jogtulajdonosoktól. A PANDORA viszont csak az archivált anyag egy kis részéhez nem enged hozzáférést – olyan oldalakhoz, amelyek üzleti vagy más szempontból érzékeny tartalmúak. Ami az archiválásra való ajánlás lehetőségét illeti: a Library of Congress egyértelműen jelzi, hogy nem fogad el ilyen javaslatokat, a UK Government Web Archive és a WAX honlapján semmilyen tájékoztatás nincs erre vonatkozóan, míg néhány más archívumnál van rá lehetőség valamilyen módon.

Személyes szolgáltatások: A National Library of New Zealand katalógusában a felhasználók elmenthetik, és később újrafuttathatják a keresőkérdéseiket, és mivel a webarchívum is kereshető az OPAC-ban, ezért értelemszerűen ez a funkció rá is kiterjed, még ha nem is arra lett specializálva. Az IA Wayback Machine honlapján ugyan van regisztrálási lehetőség, de az így létrehozott felhasználói fiók csak az IA egyéb gyűjteményeibe történő tartalomfeltöltésre szolgál, úgy tűnik, hogy a webarchívumhoz nincs köze. Más archívumoknál sem sikerült személyre szabható szolgáltatásokra bukkanni. A PANDORA az egyetlen, amelynél nyilvános havi jelentés készül az újonnan archivált tételekről, de ez persze nem ugyanaz, mintha a saját érdeklődési körüknek megfelelő témafigyelést állíthatnának be a felhasználók. Ilyen funkciók már meglehetősen elterjedtek az OPAC-okban és a digitális könyvtárakban, úgyhogy érdemes volna a webarchívumokba is beépíteni hasonlóakat, hogy több, rendszeresen visszatérő felhasználójuk legyen.

Adatbányászat: Egyik archívumnál sincs említés arról, hogy valamiféle adatbányászati lehetőséget nyújtana. A UK Web Archive rendszerét nemrég két vizualizációs technikával egészítették ki: az egyik egy címkefelhő, a másik pedig egy 3D-s animált fal, de úgy tűnik, hogy ez a rendszer sem nyújt segítséget az adatbányászattal foglalkozó kutatóknak. Annak sincs sehol nyoma, hogy a webszerverek naplófájljait is archiválnák, pedig ezek a *log* állományok fontos technikai adatokat tartalmaznak (pl. operációs rendszer, böngészőverzió, sávszélesség), amelyek a webtechnológia fejlődésének kutatásához nagyon hasznosak lennének. Természetesen az eredeti szolgáltatók saját célokra egy ideig általában megőrzik ezeket, de hosszú távú archiválásuk rendszerint nem megoldott.

Webhely rekonstruálás: A webarchívumok elvileg alkalmasak lennének arra, hogy legalább részben helyre lehessen állítani belőlük véletlenül elveszett vagy szándékosan törölt site-okat. A *Frank McCown* által készített WARRICK segédprogrammal jó esetben visszanyerhető egy webhely tartalma az IA Wayback Machine gyűjteményéből, illetve a Google, az MSN és a Yahoo keresőjének *cache* tárolójából. Azonban a jelen kutatásban vizsgált tíz archívum egyikénél sincs erre a lehetőségre utaló információ, vagyis úgy tűnik, hogy egyik sem nyújt ilyen szolgáltatást. Mivel a WAX legalább 3 hónapos, az IA Wayback Machine pedig 6-12 hónapos késéssel teszi nyilvánossá az archivált anyagot, ezeknél azonnali helyreállításra nem is lenne mód.

Keresőgépekkel való indexelhetőség: Egyes archívumok kifejezetten kitiltják a keresőgépek robotjait, míg mások csak a kezdőlapot vagy csak a metaadatokat tartalmazó lapokat engedik leindexelni. Utóbbira jó példa a UK Web Archive és a PANDORA, mert ezek megjelennek a Google találati listáiban, de ha rájuk kattintunk, akkor nem az archivált weboldalon találjuk magunkat, hanem az oldalhoz tartozó információs lapon, és innen még egy kattintás kell az archív példány eléréséhez.

Nem archivált tartalom kezelése: Ha egy olyan URL-t keresünk, amely nem szerepel az archívumban, akkor mindegyik rendszer visszaad vala-

milyen hibaüzenetet. Ebben benne van a kért URL cím, valamint annak a magyarázata, hogy ez miért nincs meg az archívumban, és hogy milyen alternatív lehetőségeink vannak. Ha az URL cím mögött van élő weblap, akkor az IA Wayback Machine automatikusan lementi azt és erről értesíti a felhasználót egy *banner* csíkon az oldal tetején.

Összességében elmondható, hogy bár a vizsgált tíz angol nyelvű webarchívum többségében rendelkezik az alapfunkciókkal (pl. URL és kulcsszó szerinti keresés, szűkítési opciók), a fejlettebb lehetőségek (pl. adatbányászat, személyre szabás, site-helyreállítás) mindenütt hiányoznak. Használhatósági problémák is vannak még némelyiknél (pl. eldugott súgó, nehezen megtalálható összetett kereső). Valószínűsíthető, hogy ezek a hibák és hiányosságok még olyan gyermekbetegségek – különösen a csak néhány éve indult archívumoknál –, amelyeket idővel majd kijavítanak, illetve. pótolnak a fejlesztők, mivel eddig inkább a rendszer felállítására és a gyűjteményépítésre fordították az erőforrásokat. Vannak is erre utaló jelek, mert négy hónappal a cikkben ismertetett kutatás után már néhány hasznos újdonságot fel lehetett fedezni: a UK Web Archive például megjelenít egy *n-gram* grafikont, amely azt mutatja, hogy hogyan változott időben az adott keresőkérdés gyakorisága; a UK Government Web Archive pedig a találati eredményeket már témák szerint klaszterezni is tudja, valamint megengedi a tematikus szűrést a keresésnél.

Hivatkozások

[1] International Internet Preservation Consortium Access Working Group: Use cases for access to Internet Archives. International Internet Preservation Consortium. 2006.

[2] RAS, M. – BUSSEL, S. V.: Web archiving user survey. 2007. július

[3] COSTA, M. – SILVA, M. J.: Understanding the information needs of Web archive users. 10th International Web Archiving Workshop, Vienna. 2010. szeptember

/NIU, Jinfang: Functionalities of Web Archives. = D-Lib Magazine, 18. köt. 3–4. sz. 2012./

(Drótos László)

Zenei anyagok könyvtárközi kölcsönzésének meghatározó szerepe az Egyesült Királyság zenei életében

A cikk legfőbb megállapítása az, hogy az Egyesült Királyság könyvtárközi kölcsönzésében egyedül a zenei anyagok kölcsönzése növekszik. A szerzők azt elemzik, hogy miként alakult ki ez a szerencsés helyzet és melyek ennek legfőbb jellemzői. Azt javasolják, hogy a döntéshozók figyeljenek jobban oda a további fejlesztésekre, mivel a kölcsönzés komoly szerepet játszik az ország zenei életében, ezáltal nagy hatást gyakorol milliók hétköznapjaira.

A zenei könyvtárközi kölcsönzésben különböző dokumentumok vesznek részt: zenei témájú könyvek és folyóiratcikkek, audiovizuális anyagok, önálló kották és kottalapok és messze a legnagyobb jelentőséggel az előadásokhoz szükséges *kottacsomagok*. A kották eleinte nem tartoztak a könyvtárközi kölcsönzésben gyakran előforduló dokumentumok közé, mert sokáig nem kerültek be a közös katalógus-adatbázisokba, csak a helyi opacokban voltak fellelhetők, sokáig nem volt egységes azonosítójuk (ISMN 1997-től létezik), a zenei anyagokat különleges tulajdonságaik és formátumuk miatt nehéz volt keresni és beazonosítani.

A zenei dokumentumokat (könyvek, cikkek, kották) zenekutatók, diákok, tanárok, előadóművészek, zenészek, énekesek, karmesterek kérik. De a legnagyobb felhasználói csoportra kétség kívül jellemző, hogy kottacsomagokra van szükségük: kórusok, operatársulatok, madrigál-, kamara-, templomi kórusok és természetesen iskolák. Bár a kottacsomagok megszerzésének más módjai is vannak (vásárlás, bérlés a kiadótól), de a leggyakoribb a könyvtáraktól való kölcsönzés. El lehet mondani, hogy a könyvtárak ezáltal komoly szerepet játszanak az ország zenei életének kiteljesedésében, a zenével kapcsolatos kulturális tevékenységek sok millió ember életét gazdagítják. Egy 2008-as felmérés szerint 2600 amatőr zenei együttes van 180 000 előadóval, akik 10 000 koncertet tartanak évente, 1,6 milliós közönségnek.

Hogy kerültek be a kottacsomagok a könyvtárakba? Az 1920-as években a megyei könyvtárak szolgáltatásainak kialakításánál a könyvtárak sok zenei anyagot kaptak azért, hogy a vidéki és a városi iskolák kulturális életének fejlesztésében