Műhely

Webhelyek metaadatolási problémái

ILÁCSA Szabina

Az OSZK webarchiválási pilot projektjében a Szabványosítási Iroda feladata az volt, hogy segédkezzen a webarchívum szabványos metaadat-struktúrájának kialakításában. A munka még folyik, a jelen cikk egyfajta helyzetjelentés az eddig megtett útról.

De mit is értünk a webarchívum szabványos metaadat-struktúrájának kialakítása alatt? Ahhoz, hogy erre választ adjunk, először meg kell értenünk a forrást és el kell helyeznünk az ismert bibliográfiai univerzumban. Milyen metaadatai vannak a forrásnak? Ezek hogyan aránylanak más források metaadataihoz?

A metaadatokat többféle rendszer szerint sorolhatjuk kategóriákba. Azért választottuk ehhez a munkához a National Information Standards Organization (NISO) által kidolgozott modellt, mert ez a felosztás a jogi metaadatot külön kategóriaként kezeli. A tartalomhoz kapcsolódó szerzői jogok az elektronikus kiadványoknál nagyon komoly következménnyel járnak a szolgáltathatóság tekintetében, ezért is tartottuk fontosnak ezzel a típussal külön foglalkozni. A metaadatoknak nincs egy egységes, minden szituációra alkalmazható felosztása. Amikor metaadattipizálást választunk a projektünkhöz, akkor tulajdonképpen nézőpontot választunk. Ezért fontos, hogy amikor metaadattípusokról beszélünk mindig hivatkozzuk meg, hogy melyik felosztás szerint beszélünk az adott típusról, hiszen más lehet egy típus tartalma attól függően, hogy melyik tipizálás részét képezi.

Tehát, ha a NISO-féle tipizálás szemüvegén keresztül nézünk a webarchiválásra, akkor milyen típusú metaadatokra számíthatunk ebben a közegben? Egyrészt leíró metaadatra, ami a forrás visszakereshetőségét és azonosíthatóságát biztosítja. Illetve adminisztratív metaadatra, ami tulajdonképpen egy ernyőkifejezés, amibe beleértjük azokat az információkat, amelyek a forrás kezeléséhez szükségesek, vagy amelyek a forrás keletkezéséhez köthetőek. Az adminisztratív adatok körén belül külön kiemelendők: a technikai metaadatok, információk a digitális fájlokról, amik ahhoz szükségesek, hogy dekódolni és tulajdonképpen használni tudjuk ezeket a fájlokat, megőrzési metaadat, ami a digitális fájlok hosszú távú kezelését, esetleges jövőbeni migrációját segíti és jogi metaadat, ami a tartalomhoz kapcsolódó szellemi tulajdoni jogokat írja le.[1]

Metaadat típusa	Jellemző tulajdonságok	Elsődleges használat
Leíró metaaadat	cím szerző tárgy műfaj megjelenési dátum	*discovery*[2] megjelenítés interoperabilitás
Technikai metaadat	fájltípus fájlméret létrehozás dátuma/ideje tömörítési séma	interoperabilitás digitálisobjektum-kezelés megőrzés
Megőrzési metaadat	checksum megőrzési esemény	interoperabilitás digitálisobjektum-kezelés megőrzés
Jogi metaadat	szerzői jogi státusz licencfeltételek jogtulajdonos	interoperabilitás digitálisobjektum-kezelés
Strukturális metaadat	sorrend hely a hierarchiában	navigáció
Jelölőnyelvek	bekezdés heading lista név dátum	navigáció interoperabilitás

1. ábra
A NISO metaadat-felosztása tipikus jellemzőkkel[3]

Eltérő mértékben ugyan, de minden forrás többféle típusú metaadattal rendelkezik. Könyvtári környezetben jellemzően olyan forrásokkal dolgozunk, amelyeknél leghangsúlyosabban a leíró metaadatok vannak jelen és a többi kategória adatelemei (már ha vannak) nem annyira hangsúlyosak, illetve használói érdeklődésre sem feltétlenül tartanak számot ezek az adatok. A webarchívumnál a leíró metaadat mellett meghatározó az adminisztratív és ezen belül/mellett főleg a technikai metaadat, illetve, ahogy azt már fentebb említettem, a jogi metaadat. Sőt, bizonyos esetekben ezek jelentősebbek is, mint a leíró adat. Használói érdeklődésre is számot tartó információ például az is, hogy milyen mélységű a mentés, vagy van-e valami, amit nem sikerült lementeni. Kutatási szempontú webarchívum-használatnál fontos tudnia a kutatónak, hogy a mentés mennyire ad hű képet a forrás akkori állapotáról. A tartalomhoz kapcsolódó szerzői jog tulajdonosának hozzájárulása nélkül nincs lehetőség az archivált webhely nyilvános szolgáltatására, így a forrás jogi helyzetéhez kapcsolódó adatok megfelelő nyilvántartása a webarchívum mint szolgáltatás alapját képezik.

Tehát amikor egy webarchívum metaadatstruktúráját tervezzük, akkor ezekre a típusú metaadatokra is gondolnunk kell.

Metaadatok a levéltárakban és a könyvtárakban

A webarchívumok metaadatolására nincsenek nemzetközileg kiforrott jó gyakorlatok,

– egyrészt mivel ez a bibliográfiai forrás (továbbiakban forrás) közgyűjtemények keretei között meglehetősen újnak számít,
– másrészt mivel két, nagyon különböző metaadatolási szokásokkal rendelkező intézmény végez webarchiválást: a levéltár és a könyvtár.

Kiforrott jó gyakorlat ugyan még nincs, de 2018-ban megjelent egy ajánlás az OCLC részéről[4]. Az ajánlás hat célt fogalmazott meg:

1. Kialakítani közösség- és szabványesemleges gyakorlatokat az archivált webtartalmak leíró metaadatainak számára, tekintettel mind a végfelhasználók, mind a metaadatolással foglalkozó gyakorló szakemberek igényeire.
2. Definiálni egy alap adatelem-csoportot használati

útmutatásokkal, amik az adattartalom létrehozására vonatkoznak.
3. Biztosítani azt, hogy a megfogalmazott adatelemek használhatóak legyenek más szabványokkal/szabályzatokkal együtt is, amely szabványok/szabályzatok sokkal részletezőbb adatelem-készlettel rendelkeznek.
4. Hidat képezni a bibliográfiai és a levéltári megközelítés között a leírás tekintetében.
5. Rugalmas megközelítést kell használni, hogy ne legyen szükség túl részletes leírásra, vagy hogy ne legyen szükség a rekord alapjait érintő változtatásra az idők során.
6. Megnyugtatni a gyakorló szakembereket, hogy bár a terület még fejlődőben van, amit csinálnak, az egy következetes gyakorlattá fog összeállni.

Amint az látható, az OCLC ajánlás egyfajta közös alapot kívánt teremteni a már létező gyakorlatok között. De mitől lettek ennyire különbözőek ezek a gyakorlatok?

Mint már említettem, a levéltári és a könyvtári terület nagyon különböző metaadatolási szokásokkal rendelkezik. Hogy miben áll ez a különbség, az egy meglehetősen komplex kérdés, mivel a gyakorlatban az állomány eseti sajátosságai miatt a helyzet nem olyan fekete-fehér, mint ahogy azt most ábrázolni fogom. Ugyanakkor, mivel nem ez a cikkem fő témája, indokoltnak tartom a kérdés leegyszerűsítését. Aki bővebben szeretne az egyes intézmények metaadatolási szokásairól olvasni, annak ajánlom kiindulópontként a *Levéltári kézikönyv* és a *Könyvtárosok kézikönyve* megfelelő szakaszait.[5, 6]

A levéltári és a könyvtári feldolgozás közti alapvető különbséget az általuk gyűjtött anyag jellege adja, a metaadatolás csupán leképezi ezt a különbséget.

A levéltári anyag iratszinten jellemzően önmagában nem, vagy csak nagyon nehezen értelmezhető, nem publikált egységekből áll. A kontextus elsődleges fontosságú, mind a feltáró munka, mind a felhasználás során. A levéltárak alapvetően csoportokban gondolkodnak az állományukról. Az őrizetükben lévő levéltári anyagot kisebb-nagyobb részekre tagolják, és az így kialakított részeket csoportosítják, kialakítják azok egymáshoz való viszonyát, végül meghatározzák a helyüket a teljes levéltári anyagon belül.[7] Az eligazodást segítendő az állományról különböző segédleteket készítenek. A levéltári segédlet csak arra a levéltárra vonatkozik, ahol előállították.

A könyvtár nagy példányszámú, publikált forrásokkal dolgozik. A forráshoz kapcsolódó metaadatok nyilvánosak, a használók számára máshonnan is hozzáférhetőek. A metaadat-tartalom alapvető kiindulási alapja a forrás önleírása. A forráshoz kapcsolódó metaadatok – az előállító gazdasági érdeke miatt – a lehetőségek szerint egyediek, és alkalmasak a forrás beazonosítására. Ahol ezt az egyediséget nem sikerült megvalósítani, ott a könyvtáros a hozzáférési pont megfogalmazásakor ezt kiegészítő információk megadásával pótolja. A forrás leírása független az elhelyezéstől és jelentős részben attól is, hogy melyik intézmény birtokolja. Az intézményspecifikus információk jól elkülönítettek. Ezért a leírás alkalmas arra, hogy mások újra felhasználják.

Hogy néz ki ez a két szemlélet, ha a webarchívumokra vetítjük őket?

Ha kizárólag levéltári forrásként tekintünk az archivált oldalra, akkor:
- már eleve adunk a felhasználónak kategóriákat, az anyagot nem feltétlenül magának kell összeállítani,
- viszont, mint minden kategorizálásnál, itt is csak törekedhetünk az objektivitásra, de teljesen nem fogjuk tudni elérni. A válogatás tükrözi a válogatója tudását, elképzelését stb., van ugyan lehetőségünk a leírásban megadni, hogy mit tekintünk a kategória határának. Ettől persze a szubjektivitás nem csökken, viszont legalább támpontot adunk az értelmezéshez.
- Mivel itt nem egy darab fizikai példányt kell elhelyeznünk, egy forrás több csoportba is besorolható; így a felhasználó kissé kevésbé van korlátozva az általunk felállított kategóriák által, mint a papír alapú gyűjtemény esetén lenne, a korlátozás ugyanakkor adott.
- Ha egy konkrét webhelyet keresünk és nem tudjuk az URL-jét, akkor csak a teljes szövegű keresésben reménykedhetünk.
- Ha az egyedi forrás szintje egyáltalán nincs feltárva, az nemcsak a felhasználónak okozhat problémát, hanem nekünk is. Például ha egy új csoportot szeretnénk kialakítani, akkor a válogatást kénytelenek leszünk külső forrásra támaszkodva elvégezni.

Ha viszont kizárólag könyvtári forrásként tekintünk az archivált oldalra, akkor:
- a forrás több szempont szerint is kereshetővé válik és nem csak szubjektív kategóriák szerint,
- a felhasználót nem kötik a csoportjaink, a ke-

reséskor a találati halmazban maga alakíthat ki csoportokat,
– a felhasználó több információt kap az adott forrásról,
– nagy tömegű anyag feldolgozása szinte lehetetlen kihívás emberi erőforrás számára. Elvi lehetősége van annak, hogy automatikusan is kinyerhetünk metaadatokat a forrásból, viszont a jelenlegi tapasztalat azt mutatja, hogy még nem nyerhető ki a forrásból megfelelő minőségű metaadat. Ha a magyar webtartalom-előállítók között elterjed az archívumbarát weboldal ajánlás[8] szerinti weboldalkészítés, érdemes lesz újra megvizsgálni, hogy mennyiben automatizálható a leírás.

Szóval melyik nézőpont „A" helyes, ha a webarchívumokról beszélünk?
Ez attól függ, hogy milyen webarchívumot akarunk létrehozni. Milyen széles a gyűjtőkör? Milyen mennyiségű az anyag? Milyen céllal archiválunk? Milyen típusú archívumhasználatot várok a felhasználómtól? Egy konkrét forrást fog-e keresni, vagy inkább források egy csoportját? Esetleg mindkettőt? Ha csoportot, akkor milyen csoportosítás lenne a megfelelő? És hogyan fog hozzáférni a kívánt tartalmakhoz, ha a csoportosításunk nem fedte le az információs igényét?
Ahogy az látható, mindegyik nézőpontnak megvan a maga erőssége és gyengéje. Így célszerű abban gondolkodni, hogy hogyan tudnánk a két nézőpontot úgy érvényesíteni, hogy mindkettő előnyeit élvezhessük.

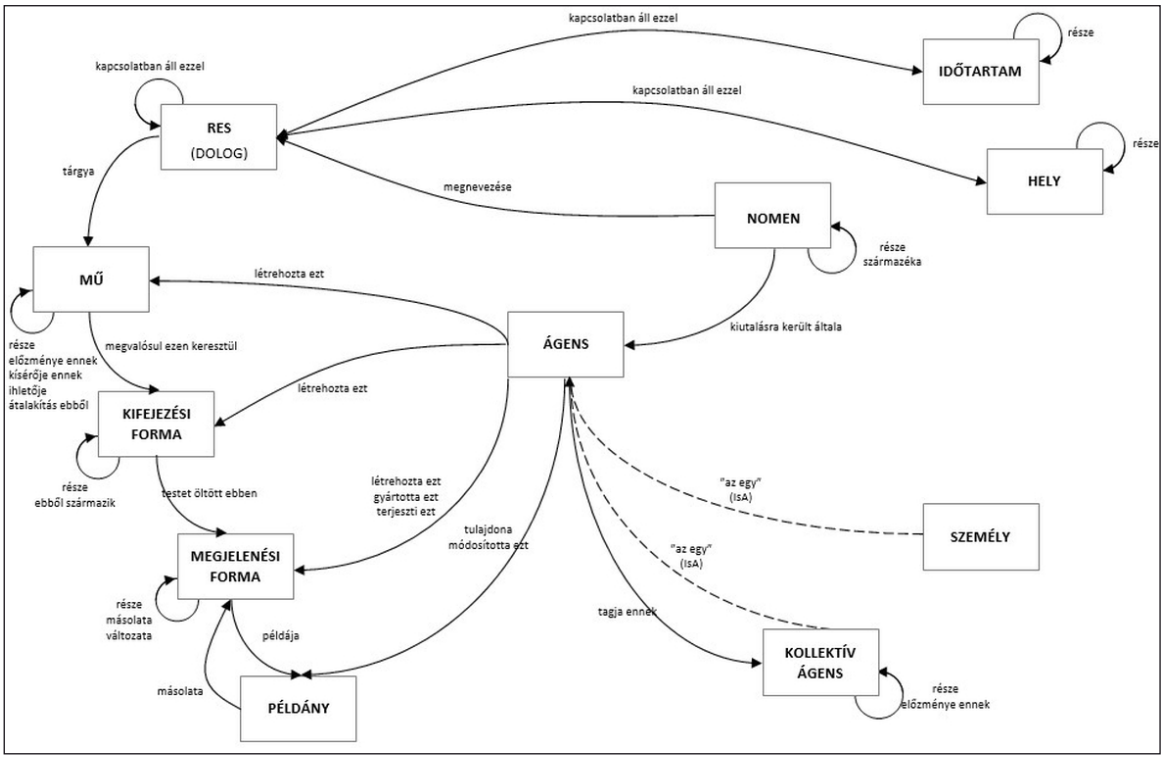
Milyen kategóriák vannak a Magyar Internet Archívumban?

Ahogy látható, a Magyar Internet Archívum (MIA) kezdetektől fogva ennek a hibrid nézőpontnak az érvényesítésére törekszik. Két kategória létezik: az egyedi leírások tárgya csak egy adott webhely, a gyűjteményi szint pedig többféle lehet a válogatás alapja szerint.

A gyűjteményi szint típusai:

– Tematikus: a válogatás alapja valamilyen témakör, szakterület, művészeti ág, műfaj, intézménytípus stb. (pl. sport, könyvkiadás, irodalom, e-periodika, egyetem);
– eseményalapú: a válogatás alapja valamilyen rendezvény, országos vagy nemzetközi szintű esemény/eseménysorozat, évforduló, katasztrófa- vagy egyéb vészhelyzet, botrány stb. (pl. világtalálkozó, olimpia, választások, centenárium, árvíz, terrorakció, tüntetés);
– intézményalapú: a válogatás köre egy adott intézmény/szervezet/cég/csoport saját online tartalmaira és/vagy a róla szóló webhelyekre/weboldalakra terjed ki (pl. az OSZK saját honlapjai és webkettes felületei);
– személyalapú: a válogatás köre egy vagy néhány személy saját online tartalmaira és/vagy a róla/róluk szóló webhelyekre/weboldalakra terjed ki (pl. Arany János művei és a vele kapcsolatos írások, képek, videók stb.);
– földrajzi hely alapú: a válogatás köre egy jelenlegi vagy múltbeli földrajzi/közigazgatási egységgel vagy annak valamely elemével kapcsolatos webhelyekre/weboldalakra terjed ki (pl. a Balaton és környéke, a Budai vár);
– vegyes: a válogatásnál valamilyen egyéb szempont volt a lényeg és ezért a részgyűjtemény nem kötődik egy adott fogalomhoz vagy dologhoz;
– webtér szintű: az aratás a magyar vonatkozású webtartalmak jelentős részére terjed ki, más válogatási szempont nincs.[9]

A gyűjteményi kategóriák között nincs formálisan kifejezett kapcsolat. Gyűjtőkör tekintetében a webtér szintű gyűjtemény mindegyik gyűjtemény felettese, de a gyakorlatban ez nem jelent rész-egész kapcsolatot. Az eseményalapú gyűjtemény több szempontból is különbözik a többi gyűjteménytől. Egyrészt a többinél jellemzően teljes webhelyeket mentenek, az eseményalapúnál pedig csak a releváns részeket, másrészt a mentés csak egy előre meghatározott, lezárt időintervallumban történik.
Jelenleg egyedi leírás arról készül, ami nyilvánosan szolgáltatható, de ez a jövőben minden bizonnyal változni fog. Az egyedi webhelyek részletes metaadatolását célszerű meghagyni a gyűjtőkör szempontjából kiemelten fontos tartalmaknak. Ezeknek a köre még nincs pontosan definiálva, emiatt azt sem tartom kizártnak, hogy a végleges változatban nemcsak minimális és részletes szint lesz meghatározva, hanem ennél több kategória kerül majd kialakításra.
A munkát az egyedi webhelyekkel kezdtük, mivel azok jobban hasonlítanak a tradicionális könyvtári forrásokhoz. Itt a könyvtári nézőpontot gondoltuk érvényesíteni, és ehhez a LRM-re átdolgozott RDA-t hívtuk segítségül.



2. ábra
LRM áttekintő diagram[11]

LRM

A Library Reference Model (LRM – könyvtári referensz modell) egy magas szinten definiált entitáskapcsolati modell, amely a tág értelemben vett bibliográfiai adatra vonatkozik. Nem tesz az adatok között olyan jellegű különbséget, hogy jellemzően milyen típusú rekordokban szoktuk ezeket tárolni. Az LRM magyar fordításban is elérhető az IFLA honlapjáról[10].

Az LRM adta nézőpont arra jó, hogy feltérképezzük vele a forrás metaadatainak belső struktúráját, illetve más forrásokkal/entitásokkal való kapcsolati hálóját. Így feltárul az adatok egymáshoz való viszonya. Az LRM egy elméleti modell, egy szemléletmód. Nem elég részletes ahhoz, hogy eszerint katalogizáljunk, viszont adhatja a vázát egy katalogizálási szabályzatnak.

Ilyen katalogizálási szabályzat például a Resource Description and Acces (RDA). Jelen munkához a már LRM szerint átdolgozott RDA-t használtuk. Az RDA az LRM koncepciójára épülő adatelemek, irányelvek és instrukciók összessége.

Ennek a cikknek nem célja az RDA részletesebb bemutatása. A téma iránt mélyebben érdeklődők figyelmébe ajánlanám az OSZK honlapján található bibliográfiát az RDA-ról eddig megjelent cikkekről.[12]

Az RDA jó alapot biztosít arra, hogy átgondoljuk, hogyan tudnánk legjobban kihasználni a metaadataink adta potenciált, majd ezt az ideális állapotot hozzá tudjuk igazítani azokhoz a korlátokhoz, amit a jelen rögzítési lehetőségeink vagy a kapacitásunk jelentenek.

Az ideális állapot kidolgozása azért fontos, mert a jelen körülmények nem jelentenek állandó helyzetet. Változhatnak a rögzítési formátumaink és változhat a kapacitásunk is. Ha már tervezéskor is csak a korlátainkban gondolkodunk, akkor lemondunk arról, hogy alakíthassuk ezt a környezetet, és az a helyzet állhat elő, hogy konzerválódik egy olyan gyakorlat, ami már meghaladható lenne.

Webarchiválás LRM-es megközelítésben

Ha az LRM szerint nézünk a webarchiválásra, akkor azt látjuk, hogy igazából két különböző forrás érdekelheti a használót. Egyrészt az az oldal, amiről a mentés készült (továbbiakban aktív), illetve maga

a mentett oldal (továbbiakban archív).
Milyen viszonyban áll egymással ez a két oldal? Ugyanannak a kifejezési formának a két külön megjelenési formájáról beszélünk vagy két külön műről? Az azonos kifejezési forma – külön megjelenési forma mellett szól, hogy a digitális másolatokat általában ennek tekintettük.
Az archivált weboldal egy fontos vonatkozásban különbözik a digitális másolatoktól. A digitális másolatok egy stabil tartalmú fizikai példányról készülnek, az archiválás viszont erősen transzformatív dolog. Adott egy folyamatosan változó forrás, ahol a tartalmak fel-és eltűnnek, hozzáadnak, meglévőt módosítanak és van, amikor levesznek tartalmakat. Archiváláskor mi önkényesen választott pillanatokban lementjük az éppen aktuális állapotot, és előfordul, hogy ez a mentés technikai okok miatt nem teljes. Lehet-e azt állítani, hogy nálunk lementett állapot hűen tükrözi az eredetije szellemi vagy művészi tartalmát? Mi úgy gondoljuk, hogy helytállóbb az, ha az aktív és az archív oldalt két külön műnek tekintjük, amelyek között szoros kapcsolat áll fent.
Ettől a kérdéstől függetlenül két leírásban gondolkodunk. A két leírást nem az a választás indokolja, hogy a két oldalt két külön műnek tekintjük. A két forrás legalább megjelenési forma szinten biztosan elkülönül, így ha a megjelenési forma és példányszintű adataikat szeretnénk jól elkülönülten kezelni, két leírásra van szükség.

Mit csináltunk eddig?

Többféle módszer létezik arra vonatkozóan, hogy hogyan lehet egy gyűjtemény metaadatstrukúráját kialakítani. Ezek a módszerek azt feltételezik, hogy a nulláról indulunk. A mi kiinduló helyzetünk azonban egy kicsit más volt. Ha valaki felkeresi a Magyar Internet Archívum (MIA) demo oldalát, akkor láthatja, hogy készültek már leírások a mentett oldalakhoz. A leírások metaadatelemeinek kidolgozása az OCLC ajánlás alapján történt. Mind az OCLC ajánlás, mind a MIA specifikációja 14 leíró elemet tartalmazott, de ezek közül csak 11 feleltethető meg egymásnak egy az egyben (*collector, contributor, creator, date, description, language, relation, rights, source of description, subject, title*)*, az *extent*et a MIA nem vette át, a *publisher* az OCLC ajánlásban nem található meg. Illetve van két elem, ami a MIA-ban bővebb értelmű, mint az OCLC ajánlásban (*genre/form* < *dc_type*; *URL* < *dc_identifier*). A 14 elem csak a leíró adatokra vonatkozik és (szinte) mindegyik tovább tagolódik alelemekre. Az egyedi webhelyekhez készült metaadatolási útmutató 91 leíró alelemet különböztet meg, az összes metaadattípus alelemeinek száma pedig 133.[13]
Ezeket a leírásokat és az ott definiált adatelemeket tekintettük a munka kiindulási alapjának. Az ottani leírások még olyan szemlélettel készültek, hogy egy leírás készül, ami vegyesen tartalmaz aktív és archív oldalra vonatkozó adatokat.
A módszerek közül a *Zeng-Quin Metadata* című könyvében található módszert választottuk tulajdonképpeni sorvezetőnek.[14] A szerzők többféle módszer és a saját tapasztalataik alapján dolgozták ki a maguk módszerét.[15, 16, 17, 18] Azért erre a módszerre esett a választás, mert az általa megfogalmazott lépések könnyebben voltak igazíthatók az elképzeléseinkhez. A módszer lépéseit bizonyos helyeken módosítottuk, vagy átugrottuk, tekintve hogy itt egy már meglévő metaadatstruktúra átstrukturálása, kiegészítése és interoperabilitásának javítása volt a cél, nem pedig egy teljesen új felépítése.

I. Szakasz – Elemezni és „megérteni" a gyűjteményt
   1. Vizsgáljuk meg a leírni kívánt forrásokat.
      Vizsgáljunk meg annyi előfordulást, amennyit csak lehet, hogy megtudjuk, mik ezeknek az információhordozóknak a közös jellemzőik.
   2. Vázoljuk fel a használni kívánt elemeket.

II. Szakasz – Elemkészlet vagy alkalmazásprofil kidolgozása a gyűjteményhez
   3. Fogalmazzuk meg a funkcionális követelményeket.
      Tegyünk fel kérdéseket: Milyen típusú felhasználókat kell kiszolgálnom? Mit akarunk kezdeni a gyűjteménnyel?
   4. Tervezzük meg a szakterületi modellt (domain model).
      Vitassuk meg és illusztráljuk: Milyen típusú dolgokat (entitásokat) fognak a metaadataink leírni? Mik a kapcsolatok ezek között a dolgok között?
   5. Azonosítsuk a kívánt metaadatelemeket a gyűjtemény számára.
      Írjuk le a
      - kívánt elemet,
      - az elem leírását és magyarázatát,

* A Dublin Core-ra épülő elemnevek. Magyarul megtalálható a MEK honlapján: https://mek.oszk.hu/html/irattar/dc.htm

- példát,
- implementációját (kötelező/opcionális? ismételhető?).

6. Döntsünk az „értékterekről”.[19]
   A fentieket egészítsük ki a következőkkel:
   - Kontrollált az érték? (igen, nem és hogyan)
   - Hogyan kontrollált? – Ez tartalmazhatja az előnyben részesített kifejezések listáját, egy létező értékszótár/séma nevét, szabályokat, ajánlást jó gyakorlatra és így tovább.
7. Keressünk létező elemeket más névterekből és „használjuk újra” őket.
   Vegyük újra elő a használni kívánt elemek listánkat, döntsük el, hogy „kölcsönzünk vagy létrehozunk”.

III. Szakasz – Előkészíteni a specifikációt

8. Véglegesítsük az elemek listáját, dokumentáljunk minden elemet egyedileg.
   Az összes elem (kölcsönzött vagy létrehozott) véglegesítve lett, szedjük őket össze egy táblázatban, ami tartalmazza a következőket:
   - elem neve (minden elemnek tartalmaznia kellene egy prefixet jelezve, hogy melyik névtérből származik),
   - elem címkéje,
   - elem URI-ja,
   - definíció (Kölcsönzött elemnél tartsuk meg az eredeti definíciót. Ehhez hozzáadhatunk egy a projektünkhöz átdolgozott definíciót is.),
   - implementálás (kötelező/opcionális? ismételhető?),
   - megjegyzés,
   - névtér kontrol (Ezt össze lehet vonni a megjegyzéssel. Adjunk egy előre definiált kifejezéslistát vagy egy létező értékszótár nevét. Vagy fektessünk le szintaxisra vonatkozó szabályokat.).
9. Készítsünk megfeleltetéseket (pl. szabványos adatcsereformátumokkal).
10. Írjunk specifikációt a teljes elemkészlethez.
11. Készítsünk iránymutatást a használathoz.
    Ezen a ponton már minden elemnél kellene lennie egy „Megjegyzésnek”, ami tartalmazza az alapvető iránymutatásokat. Ez a dokumentum a teljes iránymutatást tartalmazza.
12. Készítsünk egy XML metaadatsémát. Másik opció, hogy készítünk egy OWL ontológiát vagy fejezzük ki az elemeket RDF Schema nyelven.

A tesztelés és a revízió nincs külön lépésként feltüntetve, mivel azt a szerzők a teljes folyamat állandó elemének tekintik, még a specifikáció kidolgozása és az implementálás után is. Így a folyamat bármely pontján előfordulhat, hogy visszatérünk korábbi lépésekhez apróbb változtatásokat elvégezni.

A közös munkát tulajdonképpen a 4. lépéssel kezdtük, mivel az első hármat késznek tekinthettük. (Természetesen a fent említett tesztelés/revízió alól nem mentesültünk ez által.) A szakterületi modellünk alapját az LRM adta. Felvázoltuk az aktív és az archív oldal entitásait és ezek egymással való kapcsolatát. Erre a modellre alapozva áttértünk az 5–6. lépésre. Mivel a kiindulóalapot adó metaadatolási útmutató már eleve tartalmazott szabályozott szótárakat, logikusnak tűnt együtt tárgyalni a két kérdést. Hogy mihez érdemes szabályozott szótárat használni és mihez szabad szöveget – vagy RDA-san fogalmazva, hogy milyen rögzítési metódust használunk – annak összhangban kell lennie a 3. lépésben megfogalmazott funkcionális követelményekkel. Például ha valamilyen elemnek szűrő funkciót szánunk, akkor szabályozott szótárat kell használnunk az elem rögzítéséhez, különben az elem értékei „szétfolynak” és nem lesz alkalmas a szűrő funkció ellátására.

Az 5–6. lépés több részfeladatra bomlott. Első körben megtárgyaltuk, hogy mi az, ami leíró, és mi más típusú metaadat a leírásban. A kiinduló alap ebben az esetben is a MIA-s útmutató volt. De nem minden esetben könnyű eldönteni, hogy egy adat milyen típusú metaadat, mivel egy metaadat többféle szerepet tölthet be. Vegyük példának az aratás dátumát, ami leíró és technikai adat is egyben. Így az elemek átnézésekor volt, amit egyszerűen át tudtunk helyezni a leíró szekcióból egy másikba. Viszont ahol többféle szerepben is állhatott ugyanaz a metaadat, ott szétbontottuk az elemet aszerint, hogy milyen szerepet tölt be. A szerep szerinti külön modellezés még nem jelenti feltétlenül azt, hogy tulajdonképpen ugyanazt az elemet a leírás két külön pontján fogjuk külön-külön rögzíteni. Erre a „profiltisztázásra” pontosan azért van szükség, hogy funkció szerint külön gondolkodva az adatokról meg tudjuk állapítani, hogy hogyan „viselkednek” a különböző szerepekben, lehet-e őket egyben kezelni, vagy indokolt különválasztani. Ahogy már említettem, még nem kristályosodott ki teljesen a webarchiválás metaadatolására vonatkozó nemzetközi szintű jó gyakorlat. Ez a probléma leginkább csak az archivált tartalom leírását érinti, mivel ott fordul elő az, hogy többféle típusú intézményi gyakorlatot kellene összehangolni. Az aktív oldal le-

írására viszont már vannak kialakult szokások. Mivel ez a forrástípus jellemzően a könyvtárakat „érdekli", kevésbé problémás a leírása. Viszont ez nem jelenti azt, hogy ezen a területen ne lennének tisztázandó kérdések. Például: mint minden folytatódó forrásnak, az aktív és az archív oldalnak is van induló dátuma, az első aratás, illetve a weben való megjelenés dátuma. Megszűnéskor van lezáró dátuma, az utolsó aratás dátuma az archív esetében, de mi a helyzet az aktívval? Az-e a lezáró év, amikor a tartalom utoljára frissült vagy az, amikor az oldal többé már nem elérhető a weben? A jelen álláspontunk az, hogy az utolsó frissítés dátumát fogjuk a lezáró dátumnak tekinteni, persze bizonyos „türelmi idő" leteltével.

Ennél a lépésnél került sor a meglévő adatok aktív/archív oldalhoz való tartozásuk szerinti szétválasztására és ezen belül WEMI[20] szintekhez sorolásukra is. Ahol lehetséges volt, megtörtént az RDA elemkészletével, illetve a négy rögzítési metódussal való megfeleltetés. Ahol a MIA-n dolgozó kollegák szabályozott szótárt használtak, ott megnéztük, hogy megfelelőnek ítélnek-e egy RDA-s vagy egyéb nagy nemzetközi szótárat helyette.

Eddig csak az „alap" átnézéséről volt szó, de ezzel párhuzamosan átvizsgáltuk az RDA megfelelő szinthez tartozó egyéb elemeit is olyan szempontból, hogy használhatóak-e a webarchívum kontextusában.

Az RDA-nak való megfeleltetés korántsem volt egyszerű feladat. Ehhez először is valami „fogódzót" kellett találnunk a forráson. Milyen típusú forráshoz hasonlít leginkább az archivált weboldal? Adja magát a válasz, hogy leginkább a weboldalakhoz. Ennek azonban ellentmond, hogy a MIA-ban a webhelyek jelentős részéről nem egy mentés készül, hanem bizonyos időközönként újabb és újabb mentések, amik nem írják felül a régebbi állapotot, hanem többé-kevésbé új egységet képeznek. A „printvilág" forrásai közül így leginkább a folyóiratokhoz hasonlítanak. Tehát az archivált webartalom leírásakor hol az integráló források tipikus leírásához fordultunk analógiáért, hol a folyóiratokhoz.

Erre az analógiára azért volt szükség, hogy fel tudjuk mérni, hogy egy adott elem használható-e és ha igen, akkor mi a tartalma. Például egy folyóiratnak van gyakorisága. Van-e az archív oldalnak gyakoriság jellegű adata? Igen van: az aratás gyakorisága. Ezt az adatelemet elég volt megfeleltetni, mivel a MIA jelenleg is nyilvántartja. De továbbgondolva az analógiát, mi a helyzet a számozási adatokkal? Mivel az aratás dátumát használjuk arra, hogy a különböző mentéseket megnevezzük és sorba rendezzük, ezért tekinthetjük az aratás dátumát a részegység számozási adatának.

De van, ahol nem ilyen egyszerű a dolgunk, hiszen vannak olyan elemek is, amik kissé furcsának, mármár komikusnak hatnak webes környezetbe helyezve. Illusztrált-e egy weboldal? Ennek a kérdésnek csak akkor van értelme, ha a képi tartalmat kiegészítő tartalomként értelmezzük, – de valóban az lenne egy weboldalon? Elég sok esetben maga a képi tartalom a fő tartalom. Egy elem rögzítésének akkor van értelme, ha elmond valamit a forrásról. Ki, mikor látott utoljára olyan weboldalt, amin abszolút semmilyen képi információ nem volt? Információértéke inkább annak van, amikor ilyennel találkozunk.

A 6. lépésben az elemekhez kapcsolt szabályozott szótárakat néztük át. A vizsgált kérdések a következők voltak: Hol lehetne szabályozott szótárat használni, ahol eddig nem ezt használtak? A meglévő szótáraknál: Létezik-e olyan szabályozott szótár, ami ugyanúgy megfelel a rögzítési igényeknek, viszont jóval ismertebb/többek által használt, mint a jelenlegi? Az ez idő szerint használatban lévő elemeknél nem találtunk olyat, aminél érdemes volna a szabályozott szótár használta. Ez többnyire csak az RDA elemkészletéből használatra javasolt elemnél került elő. Ilyenek voltak például az RDA-val leggyakrabban azonosított tartalom-, média- és hordozótípus elemek és szótáraik is. Azok az esetek, amelyekben a meglévő szótár változtatását javasoltuk, leginkább olyan elemeket érintettek, ahol a hozzákapcsolt szótár „saját" volt. Vagyis tulajdonképpen a leírás tartalmazott egy előnyben részesített kifejezéslistát, amelynek elemeit nem egy már létező szabályozott szótárból választották. Ilyen volt például a *change_frequency* (a webhely változékonyságát írja le) és a *harvest_frequency* (az aratás gyakoriságát írja le) elem is. Itt a saját szabályozott szótár helyett javasoltuk a MARC 21 Publication Frequencies Scheme használatát. Ez az a szabályozott szótár, ami kódolt értékként a MARC 21 bibliográfiai formátum 008-as mezőjének 18. pozíciójára kerül. A kifejezésekhez kapcsolódó URI-k a http://id.loc.gov/vocabulary/frequencies.html oldalon érhetőek el.

Hogyan tovább?

A II. Szakaszból (elemkészlet vagy alkalmazásprofil kidolgozása a gyűjteményhez) hátra van még néhány nagyobb téma megvitatása, de rövidesen kezdődhet a tesztelés. A tesztet a MIA demo archívumában elérhető webtartalmakon gondoljuk elvégezni. A leírá-

sokat értékelni fogjuk, az esetlegesen szükséges módosításokat elvégezzük a modellünkön. Ezt követően indulhat a III. Szakasz, a specifikáció elkészítése.

Jegyzetek és irodalom

(Az elektronikus források megtekintése: 2020. január 23.)

1. RILEY, Jenn: Understanding Metadata: What is Metadata, and What is it For? Baltimore, National Information Standards Organization, 2017. 6 p. Elektronikus változat online elérhető: http://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf
2. A kifejezésnek még nem találtunk olyan magyar nyelvű fogalmat, amely az angol eredeti összes fontos jelentéstartalmát magában foglalná. A fogalom lefedi a tudatos visszakeresést és a böngészés közbeni véletlen „felfedezést" is. Az angol nyelvű verzió megtartása mellett szól az is, hogy a VuFind és a hozzá hasonló discovery- szolgáltatások is ilyen néven gyökereztek meg a magyar szakirodalomban.
3. RILEY, Jenn: Understanding Metadata: What is Metadata, and What is it For?. Baltimore, National Information Standards Organization, 2017. 7 p. Elektronikus változat online elérhető: http://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf
4. DOOLEY, Jackie – BOWERS, Kate: Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group. Dublin, OH: OCLC Research, 2018. Elektronikus változat online elérhető: https://www.oclc.org/content/dam/research/publications/2018/oclcresearch-wam-recommendations-a4.pdf
5. Levéltári kézikönyv. Szerk. Körmendy Lajos. Budapest, Osiris, MOL, 2009.
6. Könyvtárosok kézikönyve 2. Feltárás és visszakeresés. Szerk. Horváth Tibor, Papp István. Budapest, Osiris, 2003.
7. Levéltári kézikönyv. Szerk. Körmendy Lajos. Budapest, Osiris, MOL, 2009. 466 p.
8. Archive-friendly website, online elérhető: http://mekosztaly.oszk.hu/mediawiki/index.php/Archive-friendly_website
9. DRÓTOS László (szerk.): Dokumentáció. Metaadatolási útmutató részgyűjteményekhez. Utoljára módosítva: 2020.01.21. (kéziratban)
10. RIVA, Pat – LEBOEUF, Patrick – ŽUMER, Maja: IFLA könyvtári referenciamodell. A bibliográfiai információk elméleti modellje. 2017. Online elérhető: https://www.ifla.org/files/assets/cataloguing/frbrrg/ifla_lrm_2017_hun_v3.pdf
11. RIVA, Pat – LEBOEUF, Patrick – ŽUMER, Maja: IFLA könyvtári referenciamodell. A bibliográfiai információk elméleti modellje. 2017. 83 p. Online elérhető: https://www.ifla.org/files/assets/cataloguing/frbrrg/ifla_lrm_2017_hun_v3.pdf
12. http://www.oszk.hu/okr-szabvanyositas
13. DRÓTOS László – VISKY Ákos László: Dokumentáció. Metaadatolási útmutató egyedi webhelyekhez. Utoljára módosítva: 2019.08.12. (kéziratban)
14. ZENG, Marcia Lei – QUIN, Jian: Metadata. London, Facet Publishing, 2016. 216-217 p.
15. COYLE, Karen – BAKER, Tom: Guidelines for Dublin Core Application Profiles. 2009. Online elérhető: https://www.dublincore.org/specifications/dublin-core/profile-guidelines/
16. MALTA, Mariana Curado – BAPTISTA, Ana Alice: A Method for the Developement of Dublin Core Application Profiles (ME4DCAP V0.2): Detailed description. 2013. Elhangzott a DCMI International Conference on Dublin Core and Metadata Applications 2013-as Lisszaboni konferencián, online elérhető https://dcpapers.dublincore.org/pubs/article/viewFile/3674/1897
17. BAKER, Thomas – VANDENBUSSCHE, Pierre-Yves – VATANT, Bernard: Requirements for vocabulary preservation and governance. = Library Hi Tech, 31. vol. 2013. 4. no. 657-668 p.
18. ZENG, Marcia Lei – LEE, Jaesun, HAYES, Allene F.: Metadata Decisions for Digital Libraries: A Survey Report. = Journal of Library Metadata, 9. vol. 2009. 3-4. no. 173-193 p.

19. Eredetiben: value space. Az RDA terminológiában ezeket értékszótáraknak (value vocabulary) nevezzük.
20. A WEMI az LRM-ben megfogalmazott magentitások (mű = work, kifejezési forma = expression, megjelenési forma = manifestation, példány = item) angol megnevezéseinek kezdőbetűjéből alkotott betűszó.

Beérkezett: 2020. január 24.