



A digitális könyvtárak szabványai – új rövidítések

Az utóbbi három évben a hibrid és digitális könyvtárak területén jelentős változások mentek végbe. A kutatási és fejlesztési projektek helyett a hasznos rendszerek irányába történt fokozatos elmozdulás. Ennek jó példái lehetnek a Csehországban közismert digitalizálási projektek, a folyóiratokra vonatkozó *Kramerius* (<http://kramerius.nkp.cz>), vagy a történeti gyűjtemények elektronikus archiválását célul kitűző *Manuscriptorium* (<http://www.manuscriptorium.com>) és a webes tartalmak megőrzésére létrehozott *WebArchiv* (<http://webarchiv.nkp.cz>).

Manapság jellemző, hogy a kulturális, szellemi javak megőrzésére hivatott intézmények (könyvtárak, múzeumok, galériák, levéltárak) igyekeznek gyűjteményük egy részét elektronikus formában tárolni és hozzáférhetővé tenni. Adott intézmények esetenként együttműködnek például az anyaggyűjtésben, a digitális adatok leírásában, közös felhasználásában. A digitális könyvtárak építésében kihívást jelent, hogy az állományukban lévő egyes dokumentumokat nem lehet egyszerű formában, különálló elemi egységként megjeleníteni a felhasználó számára, csakis összetett módon, gyűjteményes dokumentumként. Egy digitalizált folyóirat egy tétele például tipikusan különböző képekből áll, amelyek az adott tétel egyes oldalait reprezentálják. Minden képnek tartalmaznia kell információkat arról, hogy melyik tételhez, hányadik évfolyamhoz, évhez, számhoz stb. tartozik. Tartalmaznia kell azt az információt is, hogy milyen formátumú a kép, milyen célt szolgál, mekkora a mérete, ki jogosult a használatára. Minden képet alávetnek az *OCR* (*optical character recognition* = *optikai karakterfelismerés*) módszerének, amely szöveggé alakítja az információkat, hogy az indexelt állományban a teljes szövegű keresés lehetővé váljon. Minden képet analitikus leírással látnak el, amely strukturált adatokat nyújt az egyes publikált cikkekről. Ugyanaz a cikk több oldalas is lehet, de egy oldalon több cikk is elérhető. Vagyis az összetett dokumentumoknak különböző szintjei léteznek: megadható összetett, gyűjteményes do-

kumentumként egy teljes tétel, egy évfolyam, egy konkrét év, szám vagy cikk. A digitális könyvtárnak az egyes dokumentumokhoz rendelve tárolt adatait metaadatoknak nevezzük.

Mivel a digitális könyvtárak építése és üzemeltetése költséges vállalkozás, célszerű volna, ha tulajdonosaik, működtetőik ugyanazokat a metaadatokat használnák, vagyis szabványosítanák. Ez a folyamat a maga természetes útján halad, a világon már több szabvány elfogadott, és ezek használata egyre jobban terjed. (Megjegyzendő, hogy a digitális könyvtárakkal kapcsolatos szabványosítás nemcsak a metaadatok körét, hanem a kommunikációs és applikációs protokollokat, munkafázisokat, objektumazonosítókat, formátumokat is érinti.)

A digitális könyvtárakban alkalmazott metaadatok négy csoportba sorolhatók:

- leíró metaadatok (céljuk, hogy a dokumentum könnyen elérhető legyen);
- technikai metaadatok (a dokumentum tulajdonságainak kifejezésére: pl. a szöveges dokumentum különbözik a képtípusútól);
- strukturális metaadatok (a dokumentumok és metaadatok összekapcsolására);
- adminisztratív metaadatok (a dokumentumhoz való hozzáférés, a hosszú távú megőrzés, a szerzői jogok betartásának megoldása stb.).

A hagyományos leíró metaadatok legalapvetőbb formátumai: *MARCXML*, *Dublin Core*, *TEI* és *EAD*. A *MARCXML* (*MAchine-Readable Cataloging eXtensive Markup Language*) könyvtári környezetből származik, és a bibliográfiai adatok *MARC21* (korábban *USMARC*) formátumú, XML formában történő kifejezésére szolgál. Gondozója az USA Kongresszusi Könyvtára

(<http://www.loc.gov/standards/marcxml>). A formátumot széles körben támogatják. Főképp akkor érdemes használni, amikor nyomtatott dokumentumokat bibliográfiai információkkal látunk el (pl. digitalizálásokkor).

A Dublin Core (<http://dublincore.org/>) eredetileg internetes környezetre készült, weboldalak leírására. 15 alapelemének egyszerű struktúrája lehetőséget nyújt a digitális és digitalizált dokumentumok egyszerű, tetszés szerinti leírására. Gyengéje azonban éppen egyszerűségében rejlik. Részletesebb leíráskor az adatok pontosítására ún. minősítőket lehet használni. Az egyedileg alkalmazott minősítőket regisztráltatni kell.

A TEI-t (*Text Encoding Initiative*) az a szándék hívta életre, hogy egy tetszőleges dokumentum szövegét olyan meghatározott jellemzőkkel lehessen ellátni, amelyek megkönnyítik a szöveg későbbi, hatékonyabb feldolgozását. Gondozója a TEI Konzorcium (<http://www.tei-c.org>), jelenlegi verziója a TEI P4, készül a TEI P5-ös verzió, amely tartalmazza a középkori kéziratok és ősnymtatványok digitális szabványosításának eredményeit is (l. MASTER-formátum).

Az EAD (*Encoding Archival Description*) formátum főként a levéltári anyagok nyilvántartására szolgál. A levéltári gyűjtemény egyes részeinek hierarchikus leírását teszi lehetővé. Az egyes részek leírásához MARC formátumot használnak. Gondozója a Kongresszusi Könyvtár (<http://www.loc.gov/ead>).

Az említett formátumok bizonyos mértékig egymásba konvertálhatók, valamint különböző minták állnak rendelkezésre az egyes mezők és metaadatelemek egymásnak való megfeleltetésére. A transzformálást segítő táblázatok (MARCXML ↔ MODS, MARCXML ↔ minősítetlen Dublin Core, ONIX ↔ MARCXML stb.) a Kongresszusi Könyvtár oldalain megtalálhatók (<http://www.loc.gov/standards/marcxml>).

A metaadatok további csoportjaira (technikai, strukturális, adminisztratív) vonatkozó szabványoknál még kevésbé beszélhetünk olyan rögzített formákról, mint a leíró adatoknál. Ezek még különböző szakaszoknál tartanak: már elfogadottak, elfogadtatásra várnak, keretdokumentumként léteznek, adatszótárak éppen most készül. A technikai metaadatokat ipari szabványok alapján, dokumentumtípusonként szabványosítják. A képdokumentumok számára létezik például a Z39.87 (*Technical Metadata for Digital Still Images*) szabvány (<http://www.niso.org>). A technikai metaadatok szabványosításának másik példája a JHOVE projekt (<http://hul.harvard.edu/jhove>), amely egyidejűleg tartalmazza a dokumentum fizikai formátumának és érvényességének leírását, és ellenőrzi,

hogy a formátum megfelel-e a dokumentumtípusnak formai és tartalmi szempontból.

A strukturális metaadatok speciális kategóriát alkotnak, és a digitális könyvtárak hosszú távú őrzési szempontjából a legfontosabbak. 2002-től sikertűlt világviszonylatban (nemzeti könyvtári szinten) elfogadtatni a METS (*Metadata Encoding and Transmission Standard*) szabványt, amelynek gondozója a Kongresszusi Könyvtár (<http://www.loc.gov/standards/mets>).

Az adminisztratív metaadatok különfélek lehetnek, számos könyvtárban a dokumentumok hosszú távú megőrzése szempontjából fontosnak vélt adatok leírását szolgálják (pl. hogyan keletkezett a dokumentum, hogyan digitalizálták, esett-e át konverzió, mik a technikai paraméterei, milyen alkalmazások szükségesek a használatához stb.). Ide tartozik a PREMIS (*PREservation Metadata Implementation Strategies*), szintén a Kongresszusi Könyvtár gondozásában (<http://www.loc.gov/standards/premis>).

A szabványok világába tett rövid kirándulásunk jól mutatja, hogy a megfelelő szabvány kiválasztása nem egyszerű feladat. Különösen nem az, ha összetett dokumentumokról van szó. Hogyan is működne két digitális könyvtár között az együttműködés, a tartalommegosztás, ha mindkettő más adatcsere-formátumot alkalmaz. Ezért van nagy jelentőségük a strukturális metaadatoknak, amelyek az összetett digitális vagy digitalizált dokumentumokat úgy csomagolják be egy metaadat-konténerbe, hogy azok így könnyen közvetíthetők, szállíthatók legyenek. Jelenleg legkevesebb két ilyen formátum áll rendelkezésre:

- METS,
- MPEG21 DIDL (*Digital Item Declaration Language*, ISO/IEC 21000-2:2003)

A METS tipikusan a következő hat metaadat-csoportból áll:

1. fej (<metsHdr>),
2. leíró metaadat (<dmdSec>),
3. adminisztratív metaadat (<amdSec>),
4. objektum csoport (<fileSec>),
5. strukturális térkép (<structMap>),
6. tartalmi viselkedés (<behaviorSec>).

A fej a METS-dokumentum azonosítására és kezelésére vonatkozó adatokat tartalmazza. A második és a harmadik csoport lehetővé teszi a dokumentumba (leíró, technikai és adminisztratív) metaadatok elhelyezését. A negyedik csoport

azoknak a fizikai csoportoknak a leírására szolgál, amelyekbe azok az összetevő elemek (részdokumentumok) tartoznak, amelyekből az összetett (gyűjteményes) dokumentum felépül. A METS legfontosabb része az ötödik csoport, amely egy strukturális térkép formájában írja le, hogy a 2–4. csoportokban található elemek, reszdokumentumok milyen módon kapcsolódnak egymáshoz, és milyen jelentésük van (pl. az adat a dokumentum több manifesztációjára vonatkozik-e vagy egyetlen fizikai egységre). Fontos, hogy innen lehet utalni

más METS-dokumentumokra is, ami megkönnyíti a dokumentumok összetettségének, gyűjteményjellegének kifejezését (pl. a tétel – évfolyam – év – szám – oldal – szintek mind rendelkezhetnek a magasabb szintekre mutató utalásokkal).

/VOJNAR, Martin: Standardy digitálních knihoven – nové zkratky. = Archivy, knihovny, muzea v digitálním světě, 2005. p. 57–63./

(Prókai Margit)