



Digitális megőrzés: egy megbízható digitális adattár architektúrája és technológiája

Egy digitális adattár (repository) elektronikus tartalmak hosszú távú megőrzésére szolgál. A hagyományos tárolás és megőrzés jól kiforrott technikáival szemben a digitális megőrzésnek – amelynek területén a tárolt adatoknak csak csekély része idősebb tíz évnél – számos nyitott kérdése van. Az ilyen típusú megőrzés kételkedői előszere-ttel utalnak az ún. „linkproblémára” (ami ma elérhető a weben, nem biztos, hogy holnap is elérhető lesz), a digitálisan tárolt adatok létezésének elektromosságfüggőségére, eredetének, eredetiségének bizonytalanságára, az adatok korának bizonytalan megállapíthatóságára. A fenti problémáktól mentes, megbízható digitális adattár létrehozásához szükséges technológiákat a *Rutgers Egyetem Könyvtárában (Rutgers University Libraries = RUL)* a *FEDORA (Flexible Extensible Digital Object Repository Architecture)* alapján fejlesztett digitális adattár példáján mutatjuk be.

Meghatározások

A továbbiak megértéséhez néhány kifejezés meghatározása, illetve hasonló kifejezések megkülönböztetése szükséges.

Dokumentum: olyan (nem csak papíralapú) „beszélő dolgokat” jelent, amelyek az emberek közötti szóbeli kommunikációt rögzítik későbbi ismétlés céljára. [1]

Digitális objektumok: megkönnyítik a rögzítési, szerkesztési és másolási folyamatokat, de nem tekinthetők „beszélő dolgoknak”, mivel a megfelelő infrastruktúra nélkül a digitális tartalom elérhetetlen és érthetetlen.

A Research Libraries Group (*RLG*) szerint a digitális megőrzés a végrehajtásához szükséges tevékenységekkel definiálható:

- a metaadatokkal ellátott digitális tartalom hosszú távú megőrzése olyan módon, amely lehetővé teszi az eredeti dokumentum hasonmásának elkészítését;

- az idő múlása és a változó technológiák ellenére a folyamatos hozzáférés megoldása a tartalomhoz. [2]

A digitális objektumok a digitális megőrzési és hozzáférési folyamatok alapegységének tekinthetők, amely minden információt, a metaadatot, és magát az adatot is tartalmazza. A digitális objektum egy eszközként kezelendő, csak így marad könnyen kezelhető az adat és a hozzá tartozó információ.

Megebízható digitális adattár: az *RLG* definíciója szerint a következőket teljesíti:

- a digitális megőrzés keretrendszerét az adattár és a hozzá tartozó szabályok, szabványok és technológiai infrastruktúra alkotja;
- az adattár olyan megbízható rendszer, amelyben mind a szoftverek, mind a hardverek megfelelnek bizonyos követelményeknek.

Az *RLG* a követelményekre vonatkozó ajánlásokat is megfogalmazza, amelyeknek nem minden eleme ültethető át egyszerűen a gyakorlatba, jóllehet a könyvtárak és levéltárak általában a hagyományos megőrzés területéről átvett eljárásokra építve alakították ki őket a használói segítségével.

Megebízhatóság az adattárakban

Az egyetemi könyvtárakban a megbízhatóság a tudományosságra, a hitelességre és az időtállóságra vonatkozik, szemben például az elektronikus kereskedelem területével, ahol a megbízhatóság egészen mást jelent. Amikor például egy újságnak a digitális képét és az újságcikk szövegét külön is elérheti a használó, megbízhatóbbnak érzi az el-sárgult, gyűrött, és akár a korábbi olvasók (keze) nyomát is őrző újságlapról készült képet, mint a jobban olvasható szöveges verziót, mivel nem tudja, hogy az hogyan készült, hozzáértő szakember készítette-e, végeztek-e rajta valamilyen módosítást stb.

Az effajta munkát végzőknek számos döntést kell hozniuk a digitalizálási folyamat során, például annak meghatározására, hogy mi értékes, mi megőrzendő. Sokszor kell kompromisszumokat kötniük a gazdaságosság, a tárolási költségek stb. miatt. A megbízhatóságot növeli a használók szemében, ha a metaadat-készítők hitelesítik a munkájukat, vagyis ellátják digitális aláírásukkal.

Visszatérve az újságoldal példájára, emelt szintű szolgáltatást nyújthatunk, ha az eredetit optikai karakterfelismerő programokkal szöveges állománnyá alakítjuk, és kereshetővé tesszük. A karakterfelismerő programok gyakran tévedhetnek, a hibákat korrektúrázással javíthatjuk. A különféle programok eltérő indexelést és keresési algoritmust valósítanak meg, így más-más eredményre vezethetnek. Ebből is kitűnik, hogy nemcsak a digitális objektumokat kell megőrizni, hanem a használatukhoz szükséges technológiát is.

A digitális objektum

Minden digitális megőrzés legfontosabb tervezési feladata a digitális objektum szerkezetének leírása. Az RUL-ben a digitális objektumnak része – az újságcikkes példánál maradva – az eredeti újság-lap képe (TIFF típusú képként), és a cikkek korrektúrázatlan szöveges verziója (XML-ben). Ebben az esetben a digitális objektum a következő elemekből áll:

1. Azonosító.
2. A megjelenítéshez és szerkesztéshez szükséges objektumok.
3. Ötféle metaadat:
 - a. leíró,
 - b. technikai,
 - c. forrás,
 - d. jogok,
 - e. eredet.
4. Adatfolyam (az RLG elnevezésével bájtflow), amelynek részei:
 - a. egy METS-XML térkép, amely az újság logikai és fizikai szerkezetét írja le;
 - b. az újságcikk DjVu formátumban;
 - c. az újságcikk PDF formátumban;
 - d. az újságcikk szövege XML-ben;
 - e. az újságról készült TIFF-képek tömörítve (.tar vagy .zip tömörítést használva).

Az újságról készült TIFF-képek tárolása lehetővé teszi a megjelenítési formátumok (PDF, DjVu) újbóli, illetve az azoktól eltérő, új formátum előállítását.

Az RUL-ben minden, ami az adatokkal kapcsolatos, egy helyen, a digitális objektumban található meg. Az egységbezárás garantálja a megőrzendő adatok, valamint a megőrzéshez és megjelenítéshez szükséges információk együttes mozgathatóságát, törlését, szerkesztését. Esetünkben ezt a funkciót – az összetartozó adatok és metaadatok csoportosítását – a *FEDORA* keretrendszer teszi lehetővé.

A megbízhatóság megvalósítása

A leírt architektúrában a megbízhatóság kialakítása a következő szabványok és ingyenesen hozzáférhető szoftverek segítségével történik.

Digitális aláírás

Miután egy digitális objektum bekerült az adattárba, szükséges az integritás megőrzése, azaz a véletlenszerű vagy szándékos módosítás jelzése és az eredeti verzió visszaállítása. Egy bit megváltozása például már hozzáférhetetlenné teheti az adatot vagy valamely formáját, vagy használhatatlanná teheti a megjelenítési programot, ezért az adat integritásának megvalósítására digitálisan aláírt hash-lenyomatokat használnak.

A hash-függvény egy adott bemeneti bitsorozatra (ez lehet egy fájl is) egy rögzített hosszúságú rövid bitsorozatot állít elő, amelyet lenyomatként hívunk. A hash-függvények két fontos tulajdonsága az, hogy különböző bemenetre különböző lenyomatot kapunk, és egy lenyomat ismeretében nem állítható elő az eredeti bemenet. A több ismert hash-függvény közül a leírt architektúrában az SHA-1 (Secure Hash Algorithm) használatos úgy, hogy a digitális adat eredetijéből (a digitális objektum 4/e részéből) képzett lenyomatot a digitális aláírással ellátva a digitális objektum technikai metaadataiban tárolják (ennek felhasználását l. később). Majd az egész objektum digitálisan aláírt lenyomatát is tárolják. A háttérben egy alkalmazás rendszeresen kiszámolja az egész objektum digitálisan aláírt lenyomatát, és összehasonlítja a tárolt változattal. Eltérés esetén jelez, és a biztonsági másolatokból helyreállítják a sérült digitális objektumot.

Maradandó azonosító

Szerverek költöztetése, átnevezése miatt az internetes hivatkozások gyakran változnak, az elérhetőség megbízhatatlan, nem időtálló. A megoldást egy ún. *maradandó azonosító* (*persistent identifier* = *PID*) jelenti, amely egyértelműen azonosítja és visszakereshetővé teszi a digitális objektumot. Ilyen például az *IETF* (*Internet Engineering Task Force*) nemzetközi szervezet által kidolgozott *URN*

(Uniform Resource Name = *egységes helymegnevezés*). A gyakorlatban azonban örökké érvényes azonosítók megvalósítása akadályokba ütközhet. Az URN-t általában egy konkrét címmé, URL-lé oldják fel, vagyis az URN-hez általában egy URL tartozik. A csatolók megváltozása miatt szükség van az URN-hez tartozó URL karbantartására, amit egy megbízható, örökké működő szervezetnek kellene végeznie. Másrészt, a digitális adat egy bizonyos szoftver- és technológiai környezetben készült, így későbbi megtekintése technikai akadályokba ütközhet.

Az RUL-ben a PID előállítására két eljárás használatos: a *CNRI Handle* és az *ARK (Archival Resource Key)*. [3, 4] Mindkettő azt a PID elnevezési konvenciót követi, amely szerint a neveknek a digitális adat által használt technológiától, protokolltól függetleneknek kell lenniük. A CNRI Handle és az ARK által készített PID fontos tulajdonsága, hogy a digitális adat egy-egy megjelenési formájára mutat, és nem például magára az adatra, amely metaadatok nélkül akár értelmezhetetlen.

Egy CNRI Handle azonosító előtagból és utótagból áll. Az előtagot a *Globális CNRI Nyilvántartó (CNRI Global Registry)* osztja ki, garantálva annak egyediségét. Az utótag az RUL szabályzata szerint: „[gyűjtemény].[formátum].[egyedi azonosító az RU névteréből]”.

Az ARK azonosítókban is szerepel egy egyedi szám, melyet az ARK fenntartói, a *Kaliforniai Digitális Könyvtár (California Digital Library)* és az egyesült államokbeli *Nemzeti Orvostudományi Könyvtár (National Library of Medicine)* oszt ki. Az RUL-ben ellenőrzési célra egy „átlátszatlan” nevet is generálnak. Ez a név nem utal a gyűjteményre, vagy egyéb speciális névre, mely idővel változhat, de lehetséges átlátszó nevek előállítása is. Az RUL digitális megőrzési architektúrájában automatikusan előáll a CNRI-azonosító, amely a digitális objektum egyik metaadatában szerepel.

Verziókövetés

A FEDORA rendszer a metaadatokban naplózza és rögzíti a jogosult személy által a digitális objektumban végzett módosításokat. Ha például a leíró metaadatban (a digitális objektum 3/a része) változás következik be, akkor a régi verziót mentik, az új leíró metaadat egy verziószámot kap, és ezt a szerkesztést feltüntetik az eredeti metaadatban (a digitális objektum 3/e része). Ezek után már csak a digitálisan aláírt lenyomatot kell újra előállítani.

Eredeti vagy másolat?

A hagyományos archívumokban az *eredeti dokumentum* a nem másolt, nem szerkesztett dokumentumot jelenti, szemben a digitális világgal, ahol a hálózaton való átvitel, az adatbázisok biztonsági mentése stb. során keletkezett másolatok is eredetinek tekinthetők. Az RUL rendszerében a módosított, szerkesztett objektumokat az eredeti objektumtól időbélyegekkal (date-time stamp) és digitális aláírásokkal különböztetik meg. A digitális objektum előállításakor a METS-XML fejlécébe automatikusan bekerül egy időbélyeg, a technikai metaadatba pedig a digitális objektum 4/e részének, azaz az eredeti dokumentum tárolt változatának a digitálisan aláírt lenyomata. Az eredeti objektumban megtalálható a létrehozás dátuma, olvasható a különböző módosítások naplója, megtalálhatók a különböző verziók, és a digitális aláírás ellenőrzésével a hitelesség egyértelműen bizonyítható. A felhasználók is ellenőrizhetik egy-egy objektum eredetiségét, illetve nyomon követhetik a módosításokat, meggyőződhetnek a digitális objektumok eredetivel való azonosságáról.

Egy objektumon háromféle átalakítás végezhető:

- digitalizálás,
- a digitális változat megjelenítésekor, kinyomtatásakor történő formázás,
- a digitális változat különböző formátumokra alakítása.

Az utóbbi átalakításokat – azaz mindent, ami a tárolt digitális változatot (a digitális objektum 4/e része) érinti – *migrációs eseményeknek* nevezik. Ezeket a digitális objektum 3/e részeként feltüntetett metaadatban automatikusan naplózzák, hogy hitelesen igazolható legyen az eredeti megjelenési formájával és tartalmával való egyezőség.

Az adattárolás

A megőrzési és megjelenítési funkciókat is betöltő digitális adattárak magja az adattárolási infrastruktúra. A biztonságos megőrzés és a hozzáférés, megtekintés lehetővé tétele különböző formátumú másolatok előállításával csak a megfelelő szoftver- és hardverkörnyezettel, illetve ezek menedzselésével lehetséges.

Az RUL adattárolási rendszere a nyílt szabványú *tárolóhálózaton (Storage Area Network = SAN)* alapul. A rendszer skálázható, elméletileg akár 16 millió szerver, tárolólemez, -szalag és egyéb hardvereszköz is összekapcsolható több szerver segítségével, így a hálózati terhelés kiegyenlíthető. A

SAN hálózata 2 Gb/s-os átviteli sebességet is lehetővé tesz, ezt a hálózatot csak a SAN elemei használják, mert ez független a belső LAN hálózattól. Az összesen több terabájtos tárolókapacitású lemezeket és szalagokat egy *hierarchikus tárolás-menedzsment-szoftver* (*Hierarchical Storage Management = HSM*) kezeli: az archivált anyagokról másolatokat helyez el több helyen, a szalagokra kimásolt fájlokat letörli a lemezekről, szükség esetén a szalagokról visszamásolja őket lemezekre, és felszabadítja az elévült fájlok által foglalt helyet.

A biztonságos megőrzés aktív menedzselést igényel: biztonsági másolatok készítését, többszörösen redundáns tárolást fizikailag is különböző helyeken. Az RUL-ben úgy tervezik, hogy tovább növelik a már magas biztonsági szintet. A FEDORA-fejlesztés keretében a könyvtárak digitális adattárai közötti hálózat kiépítését és tükörszerverek üzembeállítását tervezik, amelyek nemcsak a digitális objektumok másolatát őriznék, hanem a hozzáférhetőség szintjét is emelnék.

Következtetések

Bár a digitális megőrzés területén még sok a nyitott kérdés (például hogyan lehetne eldönteni két objektumról, hogy szerkezetükben és szemantikájukban megegyeznek-e), a könyvtáraknak el kell kezdeniük digitális megőrzéssel foglalkozó rendszerük kiépítését. Nem elég azonban csak üzembe

helyezni egy ilyen rendszert, hanem a használókat is meg kell győzni annak biztonságosságáról, akár úgy, hogy maguk is képesek legyenek különböző ellenőrzéseket végezni a digitális objektumokon (digitális aláírások, hash-lenyomatok), amelyek rendkívül ritka, értékes kulturális kincsek is lehetnek.

Irodalom

1. LEVY, D. (1998). Heroic measures: Reflections on the possibility and purpose of digital preservation. = Proceedings of the Third ACM Conference on Digital Libraries, Pittsburgh, Pennsylvania, 1998. p.152–161.
 2. Research Libraries Group. Trusted digital repositories: Attributes and responsibilities. = An RLG-OCLC Report. 2002.
<http://www.rlg.org/longterm/repositories.pdf>
 3. További információk: <http://www.handle.net>
 4. KUNZE, J.–RODGERS, R.: The ARK persistent identifier scheme. 2004.
<http://www.cdlib.org/inside/diglib/ark/arkspec.pdf>
- /JANTZ, Ronald–GIARLO, Michael J.: Digital preservation: architecture and technology for trusted digital repositories. = D-Lib Magazine, 11. köt. 6. sz. 2005.
<http://www.dlib.org/dlib/june05/jantz/06jantz.html/>

(Somogyi Tamás)