Dancs Szabolcs

Digitális tartalmak hosszú távú megőrzéséről a Rosetta rendszerben

Múlt év októberében látott napvilágot az Európai Bizottság ajánlása a kulturális anyagok digitalizálásáról és online hozzáférhetőségéről, valamint a digitális megőrzésről (2011/711/EU). A dokumentum 8. pontja azt javasolja a tagállamoknak, hogy „erősítsék meg a digitális anyagok hosszú távú megőrzésére irányuló nemzeti stratégiáikat, tegyék naprakésszé a stratégiák végrehajtását célzó cselekvési terveket, és a stratégiákról, illetve cselekvési tervekről cseréljenek egymással információkat". A hosszú távú megőrzés kérdése magyar viszonylatban is előkerül, amikor a 2011. évi LX. törvény hatályba lépésével létrejövő Magyar Nemzeti Digitális Archívum és Filmintézet (MaNDA) MANDALAT névre keresztelt koncepciójában a magyar digitális kulturális örökség hozzáférhetővé tételét és hosszú távú megőrzését nevezi meg két fő feladatának. Cikkemben a probléma műszaki megoldásának egyik eszközét kívánom bemutatni az olvasónak.

Bevezető

2011 októberében került megrendezésre a varsói *Lengyel Nemzeti Könyvtár*ban az a konferencia, amelynek témái a digitalizálás munkafolyamata, a digitalizálással kapcsolatos marketingtevékenységek, valamint a hosszú távú megőrzés problémái voltak. A szervezők elsősorban a *Visegrádi Együttműködés* tagállamainak nemzeti könyvtáraiból érkező kollégák részvételére számítottak, ugyanakkor a – tág értelemben vett – régió más országaiból (Ausztria, Észtország, Grúzia, Szlovénia) is invitáltak szakembereket. Magyarországot a rendezvényen az *Országos Széchényi Könyvtár* munkatársai, *Dr. Sajó Andrea* főigazgató, *Dr. Vonderviszt Lajos* e-szolgáltatási igazgató, valamint szerény személyem képviselte. A hosszú távú megőrzés kérdésének megvitatásakor többen szóba hozták az *Ex Libris* által fejlesztett *Rosetta* rendszert, de tapasztalatokról, közelebbi információkról senki nem tudott beszámolni. Az elmondottakból annyi derült ki, hogy a szakmában kifejezetten jó hírnévnek örvendő termékről van szó.

Működő megoldás a hosszú távú megőrzés problémájára: a Rosetta rendszerről

Az *Ex Libris* és az *Új-Zélandi Nemzeti Könyvtár* által közösen fejlesztett Rosetta 2009-ben került a piacra. A rendszer magját az ISO-szabványként elfogadott (ISO 14721:2003) *Nyílt Archiválási Információs Rendszer* (Open Archival Information System = OAIS) elnevezésű referenciamodellben meghatározott hat funkcionális entitás alkotja, ezek: *befogadás, a digitális objektumok tárolása, adatkezelés, adminisztráció, a megőrzés tervezése, a hozzáférésről való gondoskodás*. A Rosetta támogatja továbbá a következő metaadatszabványokat: *Metadata Encoding and Transmission Standard* (*METS*), *Preservation Metadata: Implementation Strategies* (*PREMIS*), *Dublin Core*; valamint az *Open Archives Initiative Protocol for Metadata Harvesting* (*OAI-PMH*) adatcsere-protokollt. Moduláris felépítése a digitális objektumok teljes életciklusát lefedi, bármilyen formátumú tartalomról is legyen szó. Ami architektúráját illeti, a folyvást gyarapodó digitális gyűjtemények megőrzését és kezelését támogató, skálázható infrastruktúrával rendelkezik. Az egyes modulok és az adatbázis különböző fizikai vagy virtuális kiszolgálókra telepíthetők, de létezik „minden-az-egyben" megoldás is, amikor a modulok egyetlen szerveren foglalnak helyet. A rendszer hatékony működését növelendő a „minden-az-egyben" architektúrát szimultán módon, egyszerre több szerveren is üzemeltethetjük. Rugalmas rendszerről lévén szó, a kezdeti hardverkonfiguráció a későbbiekben a speciális feladatok (pl. vírusellenőrzés, fixity) ellátása, avagy az egyre gyarapodó digitális gyűjtemény tárolása érdekében további dedikált kiszolgálókkal, munkaállomásokkal bővíthető. A rendszer flexibilis voltát erősíti az absztrakt tárolási réteg, amelynek köszönhetően az egyes modulokhoz más-más tároló hardver rendelhető.

A Rosetta rendszerarchitektúrája és az OAIS modell

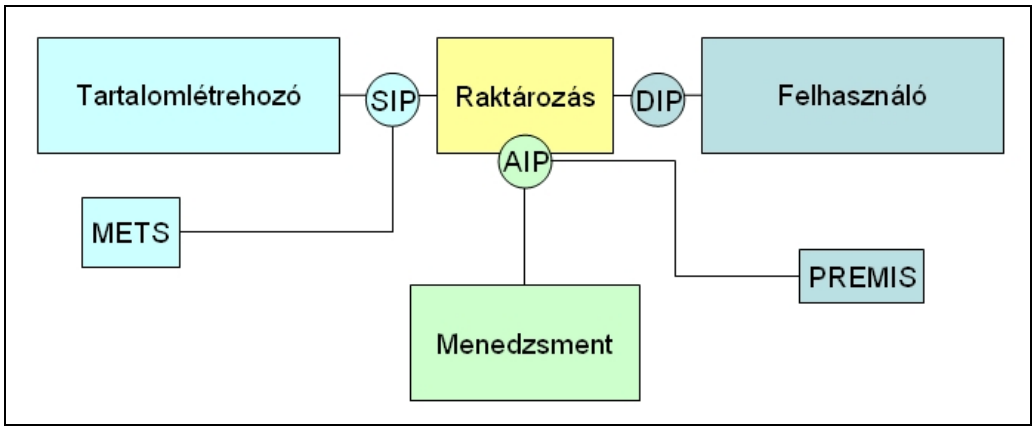
Az OAIS információs modelljének egyik alapfogalma az *információs csomag*. Egy ilyen csomag két, ún. információs objektumot tartalmaz: a *tartalmi információt* (Content Information) és a *megőrzési leíró információt* (Preservation Description Information = PDI). Maga az információs objektum egy – fizikai vagy digitális – *adatobjektum*ból és az annak jelentéssel bíró információként való értelmezhetőségét lehetővé tevő *reprezentációs információ*ból tevődik össze. A csomagokhoz további két információs objektumtípus kapcsolódhat: a *csomagolási információ* (Packaging Information) és a *csomagolási leírások* (Package Descriptions). A háromféle információs csomag:

① Submission Information Package (SIP) – a digitális tartalom előállítójától származó információs csomag;
② Archive Information Package (AIP) – az információs objektum hosszú távú megőrzéséhez szükséges információkat tartalmazó csomag;
③ Dissemination Information Package (DIP) – a felhasználónak továbbított információs csomag.

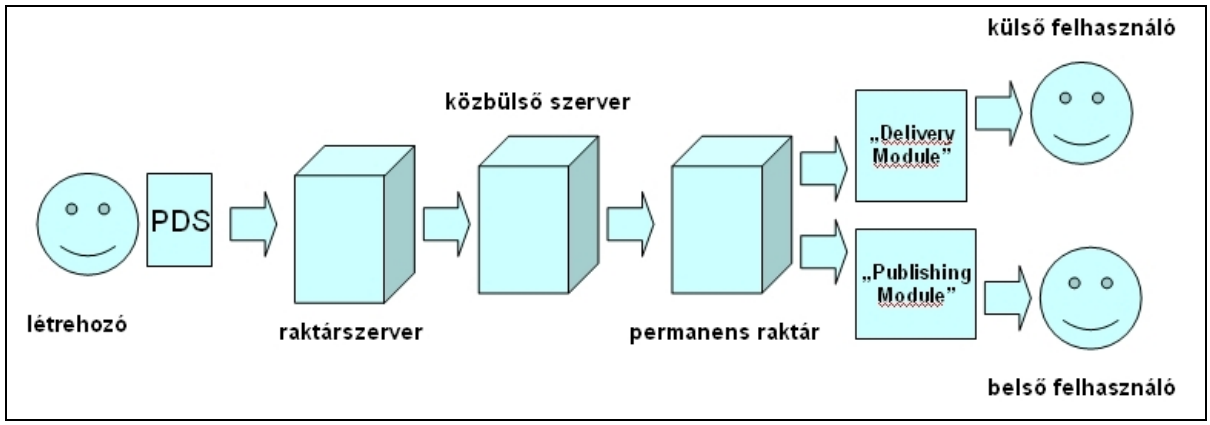
Az információs csomagoknak és a vonatkozó szabványoknak a hosszú távú megőrzés folyamatában elfoglalt helyét jól illusztrálja az 1. ábra.

Nézzük meg az OAIS modell konkrét megvalósulását a Rosetta architektúrájában!

A Rosetta webalkalmazás, amely elérhető a mai elterjedt Windows, Macintosh OS és bizonyos Linux böngészőkkel, mint pl. az Internet Explorer, Firefox, Safari vagy Opera. A felhasználói azonosítás (autentikáció) a rendszeradminisztrátor által konfigurált ún. *Patron Directory Service* (PDS) segítségével történik. A rendszer elemei közötti információáramlás útját a 2. ábra mutatja.



1. ábra **Az egyes információs csomagok helye az OAIS referenciamodellben**



2. ábra **A Rosetta architektúrája**

Mint látjuk, a digitális tartalom létrehozója a PDS-en történő azonosítás után feltölti (*3. ábra*) az adatállományokat és a rájuk vonatkozó leíró információkat (cím, szerző, létrehozás dátuma stb.) a *raktárszerver*re (Deposit Server), ahol ezek ún. *raktározási tevékenységek*ként (deposit activities) tárolódnak. Ilyen raktározási tevékenységek: a feltöltő által létrehozott, nem véglegesített tartalmak, vagyis vázlatok, piszkozatok; a digitális gyűjteményt gondozó munkatársak (staff users) által a tartalom-létrehozóhoz visszaküldött, javításra szoruló állományok; valamint a véglegesen visszautasított feltöltések.

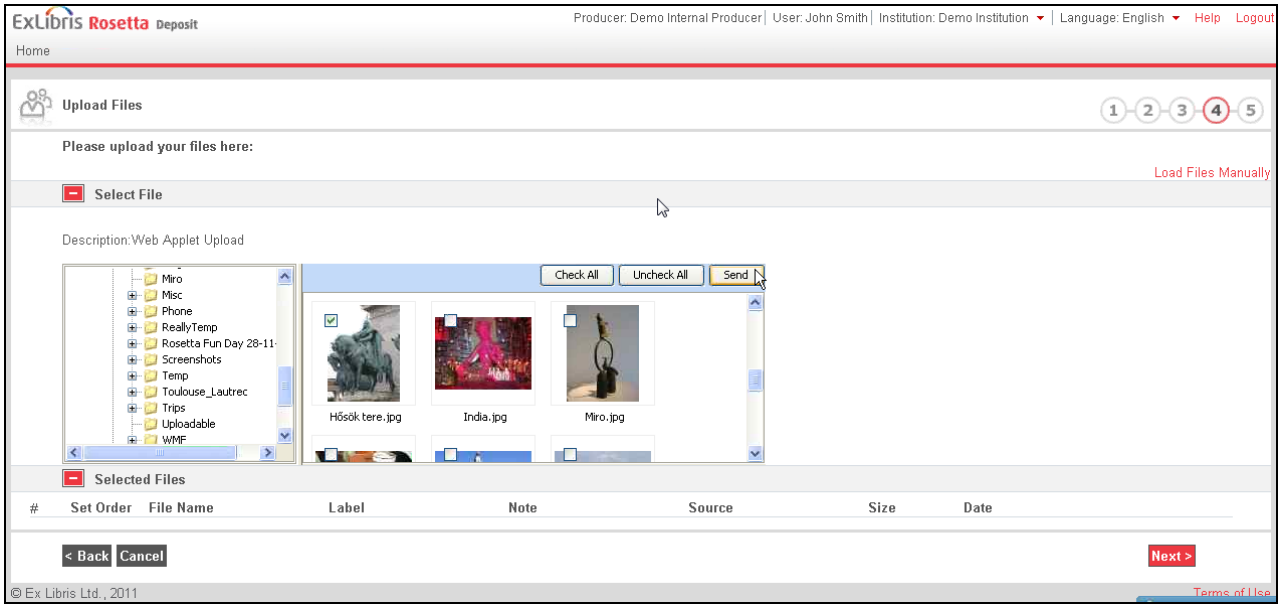
A következő állomás a *közbülső kiszolgáló* (Staging Server), ahova már SIP csomaggá konvertálva érkezik a tartalom. Az illetékes munkatársak a csomag kiértékelése után döntik el, hogy visszaküldjék, véglegesen elutasítsák, vagy tartós megőrzésre továbbítsák. A *permanens raktár*ba (Permanent Repository) ezután átkerülő, – a PREMIS terminológiáját követve – *intellektuális entitás*okként meghatározott tartalmakat nem lehet frissíteni, törölni vagy újrarendezni. Ha valamiért mégis módosítani szeretnénk valamelyiket, előbb vissza kell mozgatnunk a közbülső kiszolgálóra. A módosítást követően az entitás új verziójaként kerül eltárolásra a permanens raktárban.

A Rosetta természetesen lehetővé teszi a tartalmak megjelenítését mind a – megfelelő jogosultsággal bíró – külső felhasználók, mind a digitális gyűjtemény gondozói számára. A felhasználó egy külső alkalmazás révén küldi el kérését, amelyre a rendszer *tartalomszolgáltató modul*ja (Delivery Manager) válaszol. A felhasználói jogosultságok ellenőrzését egy ún. *hozzáférési jogosultságellenőrző* (Access Right Checker) végzi el, a digitális tartalmak megjelenítését a 4., 5., és 6. ábra mutatja.

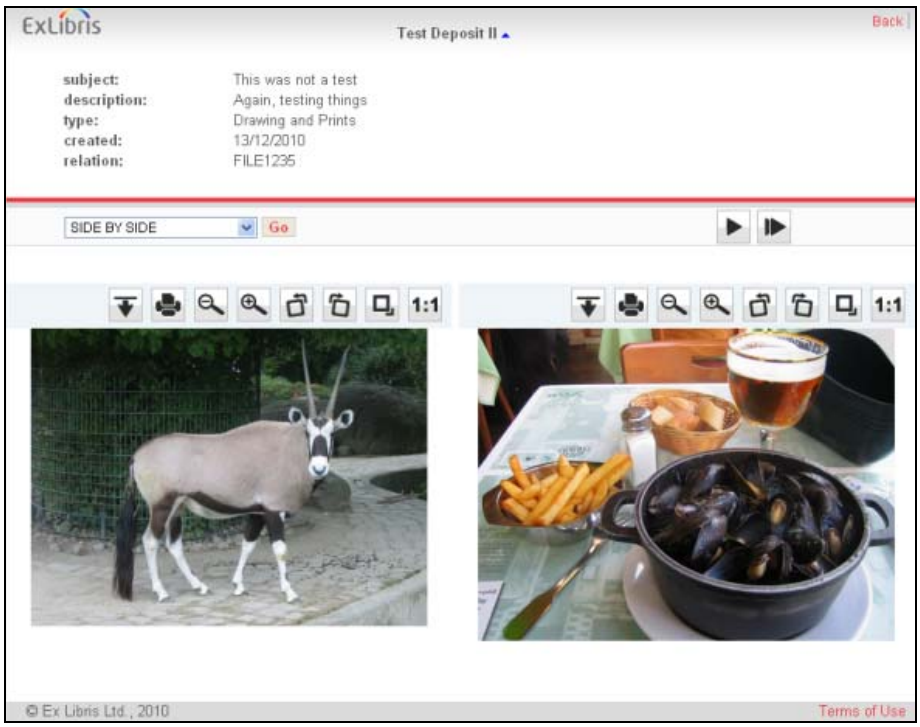
A Rosetta és a METS

A tartalomlétrehozó által végzett raktározási tevékenységek (deposit activities) adatállományokból és azok metaadataiból épülnek fel. A Rosetta a raktározási tevékenységeket intellektuális entitásokká (IE) szervezi, amelyek összetevői az adatállományok és a vonatkozó reprezentációk (az utóbbiak a digitális objektum különféle nézetei). FTP vagy NFS szervereken keresztül történő automatizált feltöltéskor a reprezentációk egy előre meghatározott tartalomstruktúra szerint szerveződnek. Ilyenkor az egyik reprezentáció állhat például bélyegképekből, míg az adatállomány egy másik reprezentációja teljes képekből.

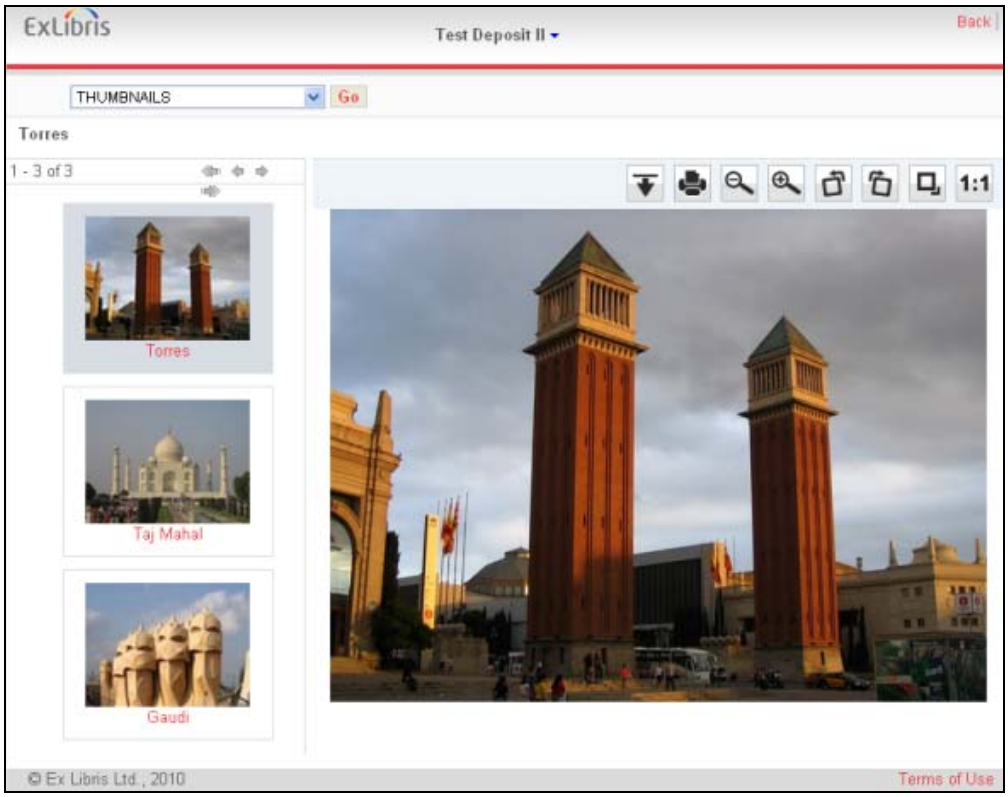
A Rosetta a tartalom létrehozója által szolgáltatott, leíró jellegű metaadatokat és a feltöltés során automatikusan generált technikai adatokat az egyes IE-hez tartozó METS-állományokká konvertálja. Az egyetlen raktározási tevékenységhez kapcsolódó intellektuális entitásokat reprezentáló METS-állományok alkotják a SIP-csomagot.



3. ábra **Digitális tartalmak feltöltése a Rosettába**



4. ábra **Digitális tartalom megjelenítése**

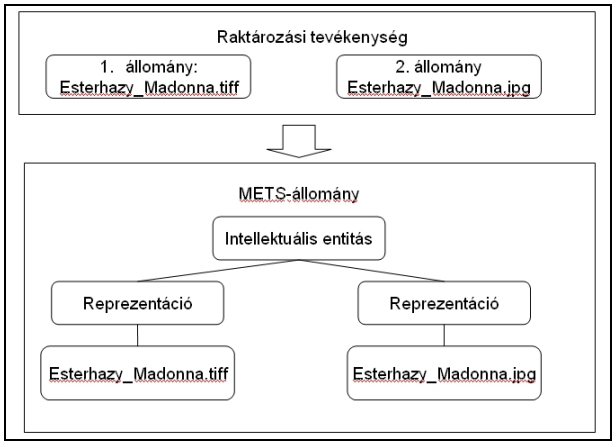


5. ábra **Digitális tartalom megjelenítése**

6. ábra **Digitális tartalmak felhasználói megjelenítései a Rosettában**

Az intellektuális entitásokra vonatkozó információkat tartalmazó METS-állományok felépítése (*7. ábra*):

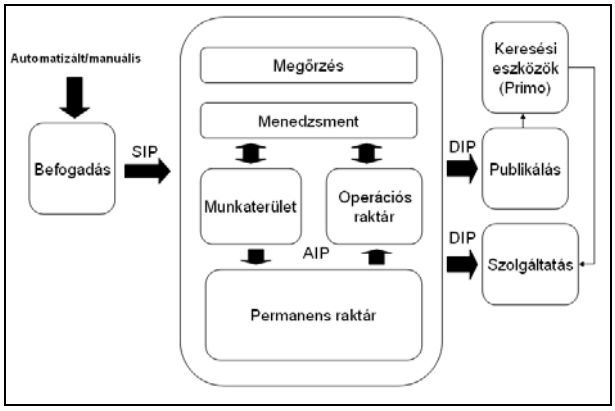
1. leíró metaadat – a tartalom létrehozója vagy a digitális gyűjtemény gondozói szolgáltatják; formátuma: tipikusan Dublin Core;
2. adminisztratív metaadat – technikai metaadat, provenienciára vonatkozó adat (pl. a feltöltő neve), hozzáférési jogosultságokra vonatkozó adat; formátuma: DPS Normalized XML (DNX);
3. struktúratérkép – az intellektuális entitások logikai csoportosításának hierarchiája.



7. ábra **Egy több reprezentációjú intellektuális entitás lehetséges példája**

A megőrzési modul

Ahogy korábban említettük, a rendszer az OAIS modellben meghatározott funkcionális entitásokra épül, ennek megfelelően kialakított moduláris felépítését illusztrálja a 8. ábra. (Ugyanitt láthatjuk az információs csomagok helyét a feldolgozás, megőrzés és nyilvánosságra hozatal folyamataiban.)



8. ábra **A Rosetta rendszer moduláris felépítése**

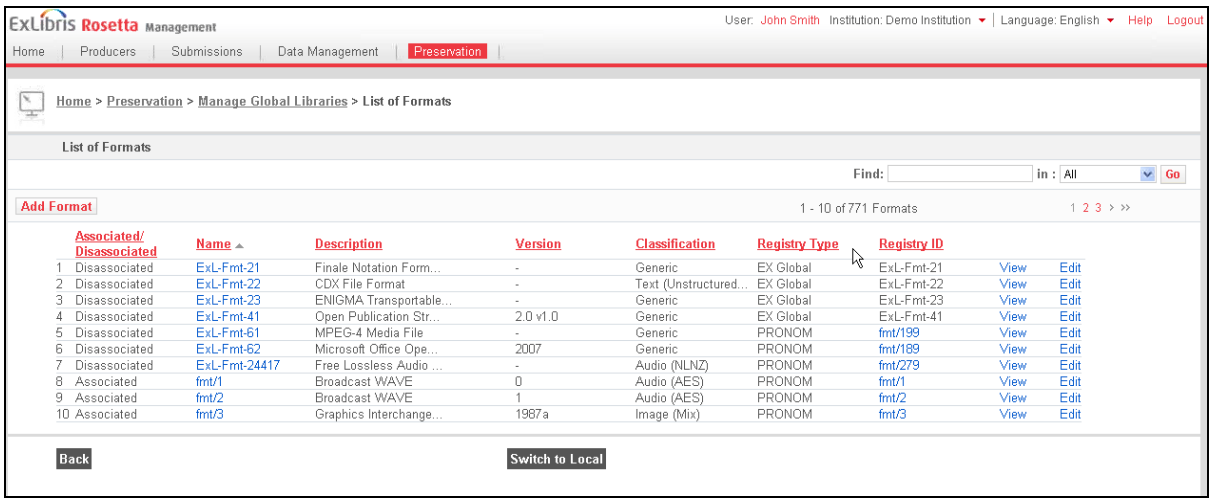
A megőrzési modul (Preservation Module) célja, hogy eszközként szolgáljon a tartós megőrzésre eltárolt digitális gyűjteményeket fenyegető lehetséges kockázati tényezők leírásához, a fenyegetett gyűjteményrészek azonosításához, a kockázati

tényezők kiiktatására vonatkozó tervek elkészítéséhez és megvalósításához.

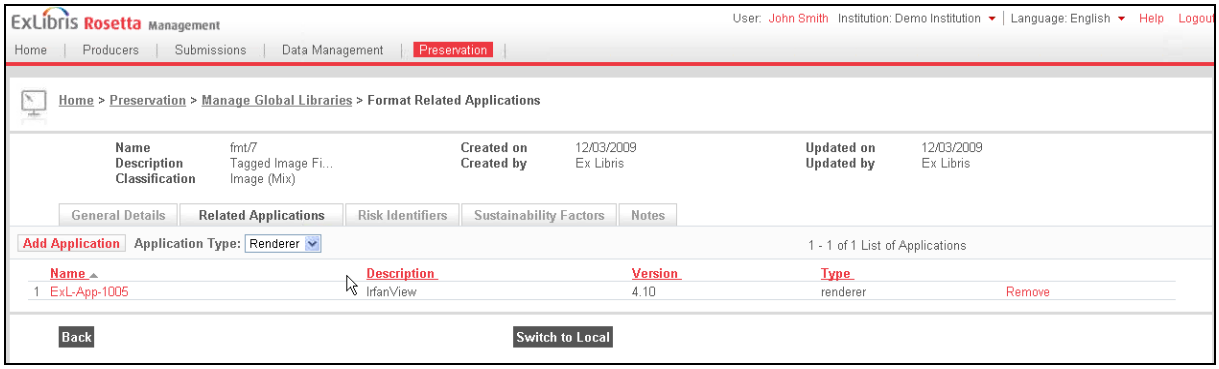
A rendszernek ez az eleme a következő almodulokból épül fel:

1. Formátumkönyvtár (Format Library) – itt a tárolt formátumokra, azok tulajdonságaira, alkalmazásaira és a velük kapcsolatos kockázati tényezőkre vonatkozó leírások találhatók, amelyeket a gyűjtemény gazdaintézményei szolgáltatnak; a Formátumkönyvtár ambíciója szerint egy globális tudásbázissá kíván válni, amelyhez a Rosetta rendszert implementáló bármely intézménynek lehet hozzáférése.
2. Kockázatelemzés (Risk Analysis) – az almodul feladata, hogy kontrollálja mindazokat az automatizált vagy manuális munkafolyamatokat, amelyek a gyűjtemény kockázati státuszának, fenyegetettségi szintjének felmérésére irányulnak. A munkafolyamatok során azonosított digitális tartalmakból létrehozott objektumhalmazokat a felhasználók továbbíthatják a Megőrzéstervezés almodulhoz.
3. Megőrzéstervezés (Preservation Planning) – az almodul szolgál azokkal az eszközökkel, amelyekre a megőrzéselemzők munkájuk során támaszkodhatnak. Segíti tehát a megőrzési tevékenységekre vonatkozó információk összegyűjtését; a szükséges tesztek végrehajtását és a teszteredmények kiértékelését; valamint, általában, a fenyegetett digitális objektumok megőrzésének érdekében történő döntéshozatalt.
4. Megőrzés-végrehajtás (Preservation Execution) – az almodul hozzárendeli a megőrzéstervezés során kreált reprezentációkat a fenyegetett intellektuális entitásokhoz. A folyamat befejeztével az intellektuális entitások új, fenyegetettségmentes reprezentációjú változatai jönnek létre.

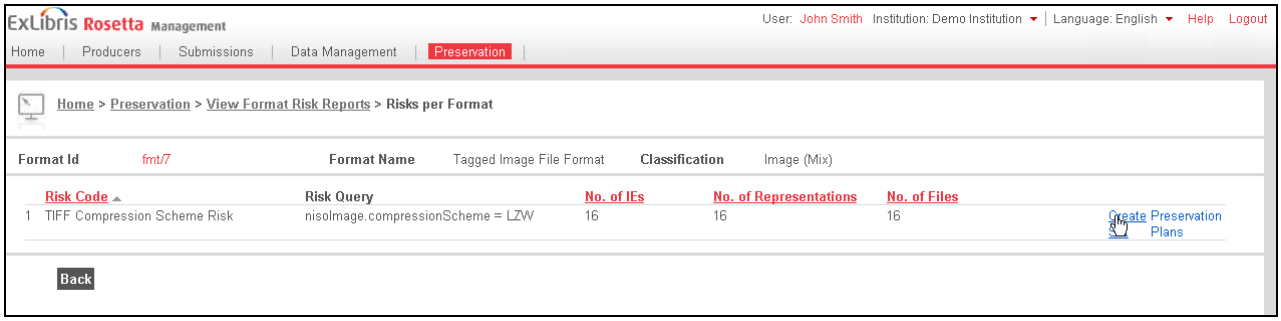
A következő illusztrációk némi betekintéssel szolgálnak a „Formátumkönyvtár” működésére vonatkozólag (*9.-13. ábra*):



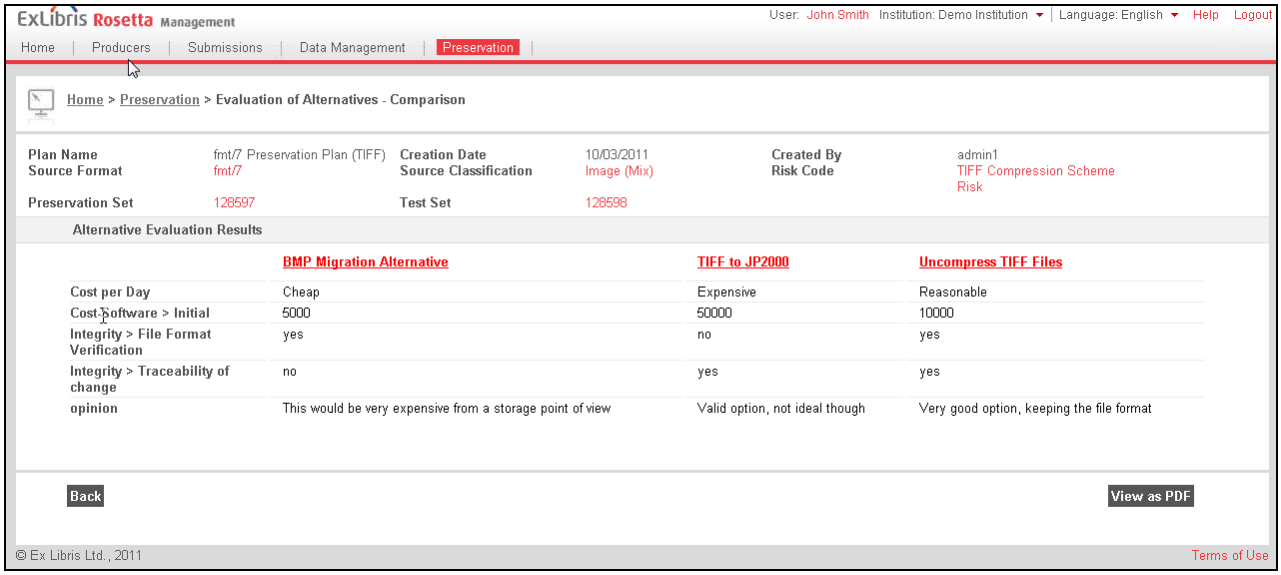
9. ábra **A formátumok listája a Formátumkönyvtárban**



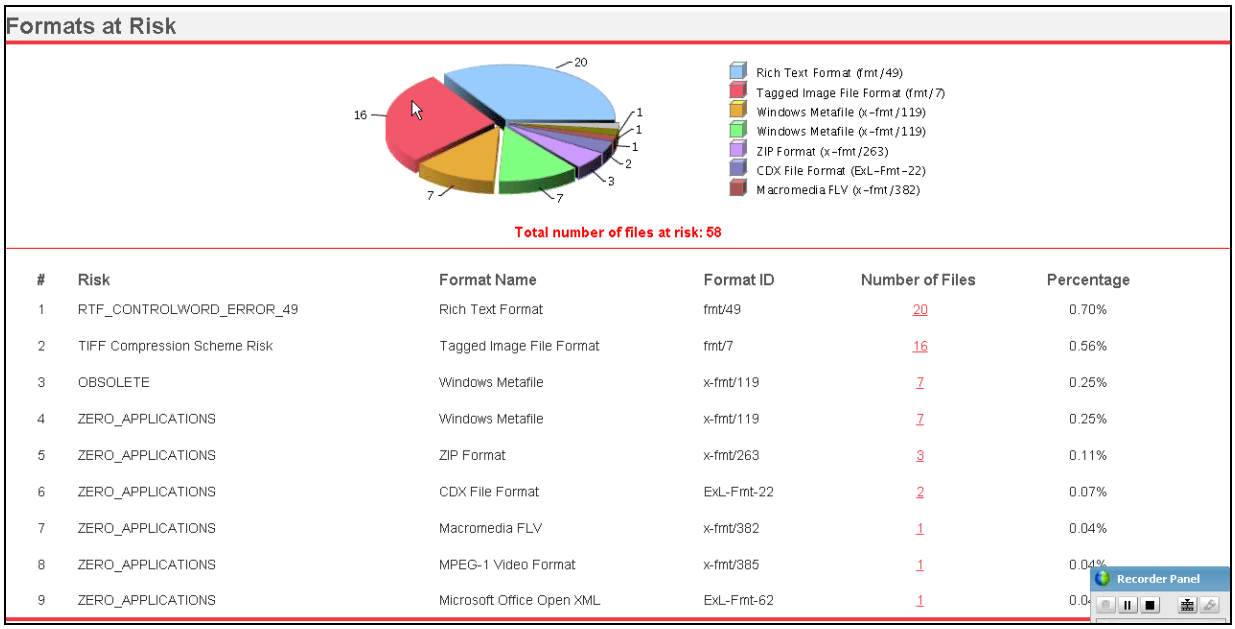
10. ábra **A formátumhoz tartozó alkalmazások**



11. ábra **Egy alkalmazással kapcsolatban észlelt kockázatok megjelenítése**



12. ábra **Megőrzéstervezés**



13. ábra **A „veszélyeztetett” formátumok kilistázása**

Mint láttuk, a rendszer működésének, egyszersmind a hosszú távú megőrzés garantálásának kulcsmomentuma a kockázatelemzés. Ennek során derül ki, hogy milyen aktuális vagy jövőbeli kockázatforrást jelenthet – példának okáért – egy formátum elavulása vagy a vonatkozó alkalmazás inkompatibilissé válása. A permanens raktárba kerülő állományok mind átesnek a kockázatelemzésen. Az analízis eredményeiből indul ki a megőrzéstervezés, amikor előbb kijelöli a fenyegetett objektumok egy teszthalmazát, meghatározza a kiértékelés során alkalmazandó ismérveket, majd alternatív módszert alakít ki a gyűjtemény szempontjából kockázatot jelentő formátumú digitális objektumok megőrzésére. A megőrzési terv tesztelését követi a megvalósítás, amely egyaránt létrejöhet belső vagy külső konverzió segítségével. Az érintett intellektuális entitások konvertálásának eredményeit a tervben megfogalmazott ismérvek alapján értékelik ki.

A Rosetta és a mormonok

A *Mormon Egyház*, teljes nevén az *Utolsó Napok Szentjeinek Jézus Krisztus Egyháza* mintegy 13 millió tagot számlál világszerte, és több mint 28 kongregációval rendelkezik. Nevükhöz fűződik a legnagyobb genealógiai szolgáltatás, a *FamilySearch* (*https://www.familysearch.org/*), amely több mint száz év aktív gyűjtésének termését foglalja magában. A 2,5 millió mikrofilmtekercsre rúgó gyűjtemény több mint 13 milliárd nevet és több milliónyi fotót tartalmaz. 2007-ben az egyház bejelentette, hogy a szélesebb körű hozzáférhetőség érdekében digitalizálja gyűjteményét. Az egyház informatikai osztályának munkatársai a digitális megőrzés biztonságos és költséghatékony eszköze után kutakodva jutottak el a Rosetta rendszerhez, és döntöttek annak tesztelése mellett.

A vizsgálat során a rendszer skálázhatóságára és befogadóképességére fektették a hangsúlyt. Még konkrétabban azt tesztelték, hogy a rendszer képes-e 24 óra alatt 200 ezer adatállomány, éves szinten tehát 2 petabyte-nyi adat befogadására, valamint horizontális particionálás esetén a Rosetta egyetlen példánya (másképpen: shardja) képes-e 50 millió rekord tárolására, amely egy húszpéldányos implementálás esetén egymilliárd rekord tárolását tenné lehetővé.

A kísérlet bebizonyította, hogy a rendszer mindkét téren eleget tesz az elvárásoknak. 200 ezer, egyenként 10 KB méretű adatállomány került feltöltésre jóval kevesebb mint 24 óra alatt, valamint a rendszer egyetlen példánya könnyedén „elbírt” 50 millió rekordot. Meggyőződve arról, hogy a rendszer kiváltképp alkalmas eszköz a jelentős méretű digitális gyűjtemények kezelésére, az egyház a Rosetta mellett tette le a voksát.

A Rosetta rendszert a Mormon Egyház mellett a következő intézmények implementálták:

Tengerentúl:
National Library of New Zealand – Új-Zéland
Archives New Zealand – Új-Zéland
National Library Board of Singapore – Szingapúr
State University of New York at Binghamton – Amerikai Egyesült Államok
National Agency for Science and Technology Information (NASATI) – Vietnam
Getty Research Institute – Amerikai Egyesült Államok

Európa:
Bayerische Staatsbibliothek (BSB) – Németország
GOPORTIS: Deutsche Zentralbibliothek fuer Wirtschaftswissenschaften (ZBW), Deutsche Zentralbibliothek fuer Medizin ZBMED, Technische Informationsbibliothek Hannover – Németország
Katholieke Universiteit Leuven – Belgium
Eidgenoessische Technische Hochschule ETH Zuerich / NEBIS – Svájc

Köszönetnyilvánítás

Köszönöm *Németh Ágoston*nak, az *Ex-Lh Kft.* ügyvezető igazgatójának, és *Ido Peled*nek, az *Ex Libris* Rosetta termékmenedzserének, hogy segítséget nyújtottak a cikk megírásához.

Irodalom

The ability to preserve a large volume of digital assets: a scaling proof of concept – *http://www.exlibrisgroup.com/files/Products/Preservation/RosettaScalingProofofConcept.pdf* (Letöltve: 2012. március 5.)

BLACKALL, Chris: Climbing Mt. Preservation: architectures and standards environments for PREMIS – *http://www.apsr.edu.au/longterm/blackall.ppt* (Letöltve: 2012. március 5.)

DAY, Michael: The OAIS Reference Model – *http://www.ukoln.ac.uk/preservation/presentations/2006/reference-models/oais-slides-day.pdf* (Letöltve: 2012. március 5.)

Reference Model for an Open Archival Information System (OAIS) – *http://public.ccsds.org/publications/archive/650x0b1.PDF* (Letöltve: 2012. március 5.)

Beérkezett: 2012. IV. 10-én.

Dancs Szabolcs
az OSZK gyűjteményszervezési igazgatója.
E-mail: *dancs.szabolcs@oszk.hu*

Jelentkezési felhívás segédkönyvtáros tanfolyamra

A Budapesti Műszaki és Gazdaságtudományi Egyetem Országos Műszaki Információs Központ és Könyvtár (BME OMIKK) emelt szintű OKJ-s segédkönyvtáros tanfolyamot hirdet.

A végzett hallgató munkaköre:
segédkönyvtáros.
Az oktatás elsősorban gyakorlati jellegű, amely a vizsgakövetelményekben is érvényesül.

A tanfolyam **2013. januárban**, keresztféléves képzési formában indul.
A képzés időtartama két félév.

A foglalkozásokat hetente egy alkalommal, csütörtökönként tartjuk, illetve minden hónap utolsó hetében kétnapos elfoglaltságot jelent a tanfolyam (csütörtök és szerda).

A tanórák mindkét napon 8 és 17 óra között zajlanak 60 perces ebédszünettel.

Részvételi díj a két félévre
150 000 Ft + a 2013-as vizsga időpontjában aktuális központi díjszabás szerinti vizsgadíj (kb. 65 000 Ft)

Felvételi vizsga nincs, a beiratkozás feltétele az érettségi bizonyítvány bemutatása.

A tanfolyam jegyzeteit, segédkönyveit kölcsönzés formájában biztosítja a szervező intézmény.

A képzésre azoknak a jelentkezését várjuk, akik a könyvtári munka gyakorlatát rövid idő alatt kívánják elsajátítani, és a számítógép használatában négy ECDL modul megismerésével jártasságot akarnak szerezni.

Jelentkezni az alábbi címre eljuttatott (kitöltött, kinyomtatott) jelentkezési űrlappal lehet:
BME OMIKK
segédkönyvtáros képzés

1111 Budapest, Budafoki út 4-6.

A jelentkezési űrlap a BME OMIKK honlapjáról letölthető
Jelentkezési határidő: **2012. december 15.**

További felvilágosítás a **463-3534**-es telefonszámon és
a **gylengyel@omikk.bme.hu** e-mail címen Lengyel Gyöngyitől kérhető.