Egy digitális könyvtár megvalósíthatósági tanulmánya

A londoni *Wellcome Library* úttörő jellegű, ötéves fejlesztésbe kezdett, melynek során egy olyan digitális könyvtárrá alakítja át magát, ahol különféle, stratégiailag fontos témákban bőséges, változatos és dinamikus tartalom érhető majd el. Az első ilyen téma a modern genetika és annak alaptudományai. A terv nemcsak méretében ambiciózus (több mint 30 millió oldalt digitalizálnak 5 év alatt), hanem abban a tekintetben is, ahogyan a dokumentumokhoz való hozzáférést és azok megjelenítését elképzelik: egyetlen „integrált könyvtárban" tárolnak majd digitalizált képeket, teljes szövegű dokumentumokat, levéltári anyagokat, videókat és hangfelvételeket, valamint eleve digitálisan született tartalmak archivált másolatait. A felhasználók mindezeket egy gazdag funkcionalitású, webkettes eszközöket is tartalmazó, vonzó felületen keresztül érhetik majd el.

Tekintve, hogy a Wellcome Library nem rendelkezik olyan infrastruktúrával, amilyennel egy ilyen szintű integrált digitális könyvtár megvalósítható lenne, ezért 2009 novembere és 2010 májusa között elkészítettek egy megvalósíthatósági tanulmányt, amelyben meghatározták a kulcskérdéseket, és hogy milyen módon lehetne elérni a kitűzött célt.

Az első kérdés az volt, hogy vajon a könyvtár meglevő, *Safety Deposit Box* (*SDB*) nevű rendszere, amelyet eddig a „born-digital", vagyis a digitális formában született anyagok megőrzésére és nyilvántartására használtak, alkalmas lenne-e nagy tömegű digitalizált tartalom menedzselésére is, beleértve olyan összetett objektumokat, mint a könyvek és a videofelvételek. Meg akarták azt is vizsgálni, hogy a JPEG 2000 képformátum használható-e ebben a környezetben, illetve hogy lehet-e „röptében" konvertálni megtekintésre alkalmas formátumokra, hogy érdemes-e átmeneti tárolót (cache) használni a megjelenítés gyorsítására, és hogy hogyan oldható meg a teljes szövegben való keresés? További két feladatként a METS metaadatszabvány alkalmazhatóságát jelölték meg, valamint egy munkafolyamatmenedzsment-rendszer beszerzésének szükségességét, mert felismerték, hogy a jelenlegi, ad hoc módon kialakított és gyűjteményspecifikus nyilvántartások hosszú távon nem lesznek alkalmasak a nagy tömegű, többféle forrásból származó, változatos formátumú digitális objektum leíró és adminisztratív metaadatainak kezelésére.

A könyvtár digitális szolgáltatásokkal foglalkozó osztálya a megvalósíthatósági tanulmány elkészítéséhez segítséget kért az SDB szállítójától: a *Tessella* cégtől, valamint a *Veridian* nevű megjelenítő rendszer fejlesztőjétől: a *Content Conversion Specialists GmbH*-tól (*CCS*).

Első lépésként a Safety Deposit Box 4-es, még fejlesztés alatt álló verzióját vizsgálták meg abból a szempontból, hogy alkalmas lenne-e a digitális könyvtár raktározási (back-end) funkcióit ellátni, valamint egy harmadik féltől származó megjelenítő (front-end) rendszerrel együttműködni? A megvalósíthatóságot demonstráló (proof-of-concept) tesztrendszer sikeresen be tudta fogadni a JPEG 2000 képfájlokat és a különféle videoállományokat is egy mintaként létrehozott beadási csomag (submission information package) formájában. Egy ilyen SIP egy XML protokoll fájlból áll, mely azt jelzi az SDB rendszernek, hogy egy objektumhalmaz várakozik betöltésre, valamint magából a tartalomból, ami egy „logikai objektum" (pl. egy könyv összes oldalképe). A fájlok formátumának beazonosításához, validálásához és jellemzőik meghatározásához készült *JHOVE* modul segítségével az SDB képes volt a JPEG 2000 állományok technikai adatait kiolvasni és ezeket az adminisztratív metaadatokat eltárolni az adatbázisában. A könyvtár által választott videoformátumokat (MPEG, WMV és Quicktime) ugyan egyelőre még nem támogatja az SDB, de a jövőben egy további fájlelemző eszközt, például a MediaInfo-t is bele lehet majd építeni. Sikeres volt a Veridian megjelenítővel való együttműködés demonstrálása is. A folyamat első lépése az, hogy a Veridian egy

submitRequest SOAP üzenetet küld az SDB-nek, és egy szolgáltatási csomagot (dissemination information package) kér tőle – az igényelt egyedi dokumentum vagy összetett logikai objektum azonosítója mellett megadva azt is, hogy az FTP szerveren belül melyik mappába kerüljenek a fájlok. A második fázisban az SDB nyugtázza a kérést, összeállítja a csomagot és felteszi a megadott helyre. Végül egy *JobCompleteRequest* SOAP választ küld a Veridiannak, melyben jelzi, hogy a kért DIP exportálása megtörtént. A Tessella további módosításokat javasolt az SDB API-jában és a beadási munkafolyamatban ahhoz, hogy minél jobban illeszkedjen a tervezett rendszerbe. A közeljövő feladata lesz az, hogy a könyvtár összevesse az SDB átalakításának költségeit és előnyeit a piacon kapható egyéb szóba jöhető *DAM* (*Digital Asset Management*), vagyis digitális vagyonkezelő rendszerek jellemzőivel, és hogy megfogalmazzon egy részletes tenderkiírást.

A digitális tartalom megjelenítésével, a teljes szövegű kereséssel és a metaadatokkal kapcsolatos kérdések esetében az ezeken a területeken jártas CCS szakembereit bízták meg a döntésekhez szükséges információk összegyűjtésével és ajánlások kidolgozásával. A sokféle kapható tartalomkezelő rendszer közül a demonstráció céljára választott Veridian a fent leírt módon képes az SDB-vel együttműködni, majd a Tessella szervereiről átvett JPEG 2000 képfájlokat JPG formátumra konvertálva – és a cache tárba töltve – azokat megjeleníteni egy webes böngésző segítségével a felhasználó számára, aki azután nagyíthatja, tologathatja, lapozhatja stb. őket. A cache beiktatása jelentősen meggyorsíthatja a folyamatot, mert ha valamelyik felhasználó egy olyan oldalt kér le éppen, amely megtalálható ebben a korlátozott méretű átmeneti tárolóban, akkor már nem kell kivárnia a konvertáláshoz szükséges időt. A cache szükség esetén kézi vezérléssel előre is feltölthető vagy kiüríthető. A felkért szakemberek csak a megvalósítandó rendszer infrastruktúrájával foglalkoztak, a leendő honlap külalakjával, a navigációs, nézegető, letöltő, illetve web 2.0-s funkciókkal nem, ezért ezek teljes körű specifikálása egy későbbi feladat lesz, mint ahogy a tartalom tényleges megjelenítési sebességének tesztelése is, ezúttal már a *Wellcome Trust* saját szervereit és háttértárait használva.

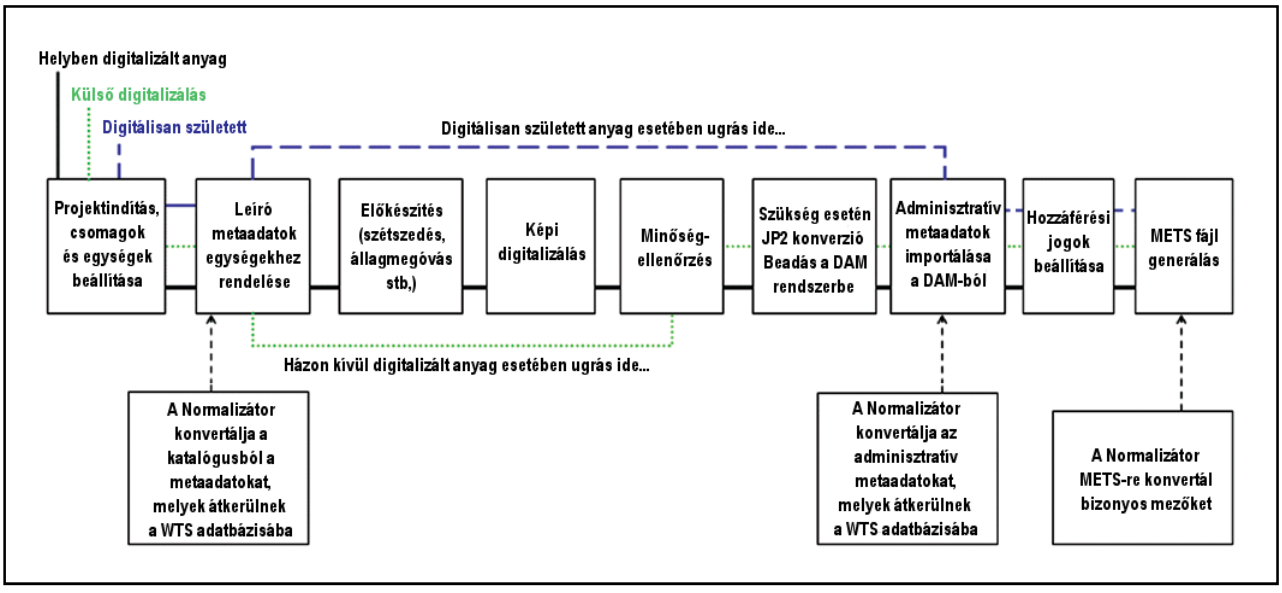
Mivel a könyvtár szeretné valamennyi digitalizált dokumentumát gépi felismertetés után kereshetővé tenni, a CCS-től erre vonatkozóan is javaslatokat kértek. A szakemberek a betűfelismerést jelentő OCR mellett a dokumentum szerkezetét felismerő ICR-t (Intelligent Structure Recognition) is ajánlották, valamint szótárak és fogalomlisták alkalmazását a fontos kulcsszavak beazonosításához és automatikus címkézéséhez, továbbá – ahol szükséges – a szöveg átírását vagy átfordítását, így maximalizálva a teljes szövegű keresés pontosságát és relevanciáját. A különböző dokumentumtípusokhoz a könyvtárnak célszerű lenne indexelési profilokat definiálni, ezáltal biztosítva azt, hogy minden esetben megfelelő módon történjék a digitalizált szöveg indexelése. Az indexelés által okozott többletterhelés várhatóan nem lesz nagy a könyvtári anyagok esetében, tekintve hogy ezek tipikusan 250-300 oldalas könyvek, oldalanként kevesebb mint 500 szóval. Átlagsebességgel számolva mintegy 150 könyvet lehet majd óránként leindexelni. Keresőrendszernek az elterjedt *Solr* szoftvert lehetne használni, amely a *Lucene* indexelőre épül, és gyorsan képes találatokat adni akár igen nagy méretű adatállományokból, akár igen sok felhasználónak is. Ha túl nagy lenne az indexállományt szolgáltató gép terhelése, akkor további szerverek is beállíthatók. Az OCR-es szövegek mellett lehetőség van átírt levéltári kéziratok, valamint eleve digitálisan született, majd archivált tartalmak leindexelésére is. A szöveges kereséssel kapcsolatos követelmények teljes körű specifikálása mellett további feladat még annak megvizsgálása, hogy a digitális gyűjtemény indexe hogyan lesz majd beépíthető a könyvtár *Encore* nevű metakeresőjébe.

A könyvtár szándéka szerint METS fájlokban tárolja majd a digitális dokumentumok metaadatait: a katalógusból származó bibliográfiai leírásokat, az SDB adminisztratív metaadatait, valamint a hozzáférést szabályozó információkat. Minden logikai objektumhoz tartozik egy ilyen fájl, amely összehozza az objektumot alkotó elemeket (pl. egy könyv oldalait, egy videofelvétel egyes szakaszait és különféle formátumait, az egy levéltári egységhez tartozó iratokat stb.). A CCS szakértői segítettek a METS fájlok szerkezetének megtervezésében és további tanácsokat adtak a használatukkal kapcsolatban. A legfontosabb javaslatuk az volt, hogy a könyvtár készítse el a saját METS profilját. Ez a profil szolgál majd hivatkozásként a könyvtár által használt METS fájlokra. Továbbá a METS *ALTO* nevű kiegészítését ajánlották az OCR-ezett dokumentumok szerkezetének leírására (vagyis az ICR adatok és a szavak helyét jelző koordináták tárolására). Mind a METS, mind pedig az ALTO széles körben használt szabványos megoldások a nagyméretű digitalizálási projektekben. A CSS

javaslata szerint a METS fájlokban a leíró adatokat a *MODS* (*Metadata Object Description Schema*) séma szerint kellene tárolni, míg a képfájlok adminisztratív metaadatait a *MIX* (*Metadata for Images in XML*) sémának megfelelően. Hogy végül is milyen metaadatszabvány(oka)t választanak a bibliográfiai és a technikai információk leírása, illetve hogy használják-e majd az ALTO kiegészítést a digitalizált dokumentumoknál, az későbbi döntésektől függ, mint ahogy az is jövőbeli feladat, hogy véglegesítsék és elkészítsék a Wellcome saját METS profilját (figyelembe véve majd a választott megjelenítő rendszer igényeit is).

Mint ahogy arról korábban szó volt, már a tervezési fázisban nyilvánvalóvá vált, hogy a könyvtárnak szüksége lesz egy munkafolyamat követő rendszerre, amellyel nyilván tudja tartani a digitalizálási műveleteket, illetve amellyel összesíteni és METS formátumban exportálni lehet a különféle metaadatokat. Ezzel a rendszerrel kellene továbbá adminisztrálni az eleve digitálisan született anyagok archiválását is. A megvalósíthatósági tanulmány készítésekor a könyvtár felvázolta egy ilyen rendszer modelljét és körülnézett a piacon a szóba jöhető termékek között. E munka összegzéseként a digitális szolgáltatások osztályán dolgozó szakemberek egy jelentésben leírták, hogy hogyan működhetne egy ilyen rendszer a gyakorlatban, hogyan illeszkedne be a digitális könyvtár architektúrájába, és hogy milyen elvárások vannak vele szemben, amelyeket majd az ajánlatok közül kiválasztott szoftvernek teljesítenie kell. Saját modelljük megtervezéséhez kezdetben a *Walesi Nemzeti Könyvtár*ban működő „testreszabott” rendszert vették alapul, amely nemcsak az egyes részfolyamatokat (pl. digitalizálás, minőségellenőrzés) tartja nyilván, hanem a metaadatokat is aggregálja, és előállítja a tartalom megjelenítéséhez szükséges METS fájlokat is. De ahogy az előkészítő munka előrehaladtával egyre jobban tisztázódott a metaadatok köre, és miután más munkafolyamat-nyilvántartásokat is megnéztek, rá kellett jönniük, hogy túl nagy elvárás egy ilyen rendszertől az, hogy képes legyen mindenhonnan begyűjteni és METS formátumban exportálni a metaadatokat. A legtöbb kapható rendszernél túlságosan sok fejlesztésre lenne ehhez szükség, és ezt a könyvtár szerette volna elkerülni. Végül is a modellt két külön rendszerből építették fel: egy *WTS* (*Workflow Tracking System*) és egy *Normalizátor* (*Normaliser*) nevű részből, melyek együttműködését a mellékelt folyamatábra mutatja *(1. ábra).* A WTS rendszer fogja nyilvántartani az egyes résztevékenységeket, és ebben csak korlátozott mértékű adatbevitel történik, míg a metaadatok nagy részének importálását, konvertálását és METS exportját a Normalizátor végzi majd.

A WTS esetében a következők az elvárások:
- projekt-, köteg- és egyedi szintű nyilvántartás;
- minden egységhez leíró metaadatok rendelése a vonalkódok alapján;
- grafikus beviteli felület a feldolgozóknak;
- felhasználók és felhasználói csoportok hozzáférési jogainak kezelése;



1. ábra **Háromféle munkafolyamat végigkövetése a WTS rendszerrel és a Normalizátorral**

- a munkafázisok követése a felhasználói input alapján (pl. kipipálható, hogy az „állagmegóvás megtörtént"), beleértve azt is, hogy mikor zajlott le valamelyik résztevékenység és ki végezte el;
- a dokumentumegységek aktuális helyének követése, szintén felhasználói bevitelre alapozva;
- parancssorból végrehajtható műveletek (mint pl. a képek JPEG 2000 formátumra való konvertálása);
- különböző jellegű, eltérő lépésekből álló munkafolyamatokhoz való rugalmas alkalmazkodás;
- a metaadatok szabványos formátumban való tárolása egy adatbázisban.

A Normalizátor, amely egy önálló, a többi könyvtári rendszertől független, de a WTS adatbázisát használó alkalmazás lesz, ilyen funkciókat lát majd el: a könyvtár katalógusából (amely a MARC21, illetve az ISAD(G) szabványokra épül) és a digitális objektumokat kezelő rendszerből átveszi a metaadatokat és leképezi a WTS által használt adatmezőkre, ily módon aggregálva őket. Ugyancsak átkonvertálja egyes kiválasztott adatmezők tartalmát a WTS adatbázisából a DAM rendszerbe (hogy a katalógusból származó metaadatok bekerüljenek oda is bizonyos adminisztrációs célokra), valamint a Wellcome saját profilja szerinti METS kimenetet is előállítja. A Normalizátort várhatóan parancssorból fogja meghívni a WTS szükség esetén. Nyitott kérdések itt is maradtak még: például, hogy milyen kész alkalmazások léteznek ezekre a leképezési feladatokra; hogy a könyvtári dolgozók képesek lesznek-e létrehozni és szerkeszteni az inputot és outputot szabályozó, XML formátumú *mapping* fájlokat; és hogy a Normalizátort valóban elég-e csak parancssorból futtatni, vagy ehhez is kellene egy önálló grafikus kezelőfelület?

A Wellcome Library azzal, hogy lefolytatott egy részletes megvalósíthatósági vizsgálatot – felmérve a meglevő rendszerei lehetőségeit és felvázolva az új rendszerekkel kapcsolatos elvárásait – képessé vált arra, hogy megtervezze és a következő két évben kifejlessze az új digitális könyvtárához szükséges infrastruktúrát. Annak köszönhetően, hogy elegendő időt szántak a modellalkotásra, az egyes részrendszerek alapvető funkcióinak és együttműködésük lehetséges módjainak specifikálására és tesztelésére, a könyvtár abban a tudatban léphet tovább a tendereztetési fázisba, hogy a főbb hiányosságokat és függőségeket sikerült felderíteni és megoldásokat találni rájuk. A tanulmány elkészítésében nagy segítséget jelentettek a külső szakértők, és a velük való konzultáció arra is felkészítette a könyvtár munkatársait, hogy hogyan kell majd kommunikálniuk a szóba jöhető rendszerek beszállítóival.

/HENSHAW, C. – SAVAGE-JONES, M. – THOMPSON, D.: A digital library feasibility study. = LIBER Quarterly, 20. köt. 1. sz. 2010. p. 53–65./

(Drótos László)