



Középkori szláv kéziratok elektronikus feldolgozása a Bolgár Tudományos Akadémián: a „Repertorium” projekt

DANCS Szabolcs

Bevezető

A középkori kéziratos források vizsgálata rengeteg tanulsággal bír mind a nyelv-, illetve irodalomtörténet, mind a filológiát, a paleográfiát segédtudományként alkalmazó történettudomány számára. A különböző kódexek, tekercsek a nyelvállapotok rögzítésén túl a szövegverziók összevetésének lehetőségével is szolgálnak, valamint – tanulmányozásuk révén – hozzájárulnak a történeti kutatások számára egyedülálló fontosságú forrásértelmezés adekvátabbá válásához. A súlyos, tudományos konzekvenciák levonásán túl az egyes kéziratok közelebbi vizsgálata olyan nem kevésbé súlyos, de az egyszerű érdeklődő figyelmét is magára irányító következtetésekhez is elvezethet, mint amilyenek a másoló jellemre vagy az általa beszélt nyelvjárásra vonatkozó megállapítások. A középkori szláv kéziratok általában egyházi, vallási témájú műveket (legendákat, szertartásszövegeket stb.) tartalmaznak. Az ezek feltárása, megőrzése érdekében kifejtett erőfeszítések a közös, európai kulturális örökség megóvásának és megismerhetővé-tételének magasstos célját is szolgálják.

A modern filológiai kutatás eszköztára, lehetőségei felbecsülhetetlen mértékben bővültek a számítástechnikának e tudományterületen való térhódításával. Szakemberek sokasága szenteli idejét, energiát újabb s újabb, az elektronikus módon történő feldolgozásnak köszönhetően – az információtudomány korszaka előtti viszonyokhoz képest – könnyen kutatható korpuszok létrehozására. A hipertext-technika lehetővé teszi az eredeti szöveg és – kritikai kiadások esetén – a kritikai apparátus közötti laza – egy kattintásnyi energiába kerülő – átjárást, a szövegváltozatok, annotációk labirintusában való gyors közlekedést, ily módon a „vízszintes elmélyülést”. Nem elhanyagolható az sem, hogy a világ egymástól távoli pontjain található forrásanyagok kerülhetnek ezáltal egy helyre, jó esetben egyenként azonos struktúrába, keretek közé illesztve, s alkothatnak egyetlen kutatható korpuszt.

Ami a szlavisztikát illeti, nem egy kezdeményezés látott napvilágot a szövegemlékek elektronikus feldolgozását illetően. Példának okáért

Cynthia M. Vakareliyska, az Egyesült Államokbeli Oregon Egyetem kutatója a 13. Szlavisztikai Kongresszuson tartott előadásában (Vakareliyska 2003) az egyházi szláv nyelven íródott középkori evangélium-szövegek elektronikus összehasonlításának megvalósíthatóvá tétele érdekében lépett fel egy jelentős méretű korpusz létrehozásának igényével. Egy a szlavisztikai kutatások számára egyelőre rendelkezésre nem álló, összehasonlító vizsgálatokra alkalmas korpusz létrehozását sürgette Andrej Bojadžiev is, a Bolgár Tudományos Akadémia munkatársa a középkori szláv szövegek oktatási célú elektronikus kiadásáról szóló előadásában (Bojadžiev 2003). (Ugyanő megjegyzi azt is, hogy más – a szláv tudományokon kívüli – területeken már léteznek hasonló korpuszok, mint amilyen például a *Thesaurus Linguae Graecae* a klasszikus tudományok terén.)

A TEI

Az elektronikus feldolgozás, illetve szövegkódolás terén betöltött vezető szerepéből következően megkerülhetetlen, hogy néhány szóban jellemezzük azt a nemzetközi kezdeményezést, a *Text Encoding Initiative*-t (továbbiakban: *TEI*), amelynek célja a területen egyfajta kváziszabvány létrehozása, a szövegek elektronikus feldolgozásának standardizálása tehát, ami érdekében a szervezet szakemberei az egyes tudományterületek képviselőivel karöltve, együttműködve ajánlásokat, útmutatókat (ún. *Guidelines*) fogalmaznak meg. Julia Flanders (Flanders 2005) a *TEI*-t egyrészt mint egy (1.) kódolási nyelvet, másrészt mint egy (2.) konzorciumot definiálja, amely utóbbinak céljai között kitüntetett helyet foglal el az említett kódolási nyelv fejlesztése. Míg a hivatalos weboldalon (<http://www.tei-c.org/>) azt olvashatjuk, hogy „a *TEI Guidelines* egy nemzetközi és interdiszciplináris szabvány”, Flanders kifejezetten azt állítja, hogy adott esetben nem egy nemzetkö-

zi, hanem inkább egy „közösségi szabványról” (community standard) beszélhetünk, amely – alkotóinak szándéka szerint – nem minden egyes dokumentumra vonatkozik. A *TEI Guidelines* alkalmazhatjuk szigorúan, de lazán is, a helyi viszonyokhoz igazítva, azaz egy rendkívül rugalmas szabványnak tekinthető. Nem különbözik ilyen értelemben az emberi nyelvtől, amely ugyancsak használható standard és nem-standard módon, s ahogy a földrajzi értelemben egymástól elkülönülő közösségek más-más dialektusokban beszélnek, úgy a *TEI Guidelines* alkalmazó különböző embercsoportok is a szabvány egyedi, rájuk jellemző használati módját fejlesztik ki. E tudományos közösségek csupán a *Guidelines* részszabályait használják fel igényeik szerint, s bővítik ki a saját, egyedi problémáik kezelését lehetővé tevő módon. A *TEI* a szövegkódolás legmegfelelőbb eszközének az ún. jelölőnyelveket, ezek közül is az *XML*-t (*eXtensible Markup Language*) tartja, amely egy szöveges sorok által megjelenített adatstruktúra, és különösen előnyére szolgál, hogy egyrészt ingyenes, másrészt alkalmazás- és platform-független. Ebből következik, hogy a *TEI Guidelines* is ezen a nyelven, pontosabban a nyelvhez tartozó séma-leíró nyelven (*Document Type Definition*, *DTD*) kerül megfogalmazásra. (A *DTD*-t tekinthetjük az elektronikusan feldolgozott dokumentumok nyelvtanának.) Az *XML* abban is segítségünkre van, hogy a szöveg ne csupán karaktersorozatként kerüljön kódolásra, hanem – példának okáért – az eredeti tipográfiairól is elképzeléssel bírjunk. Ahogy Lou Burnard arra felhívja a figyelmünket (Burnard 2005), a szöveg az olvasók képzetében felidéződő absztrakciókat jelenti, s a jelölő nyelvek célja ezen absztrakciók kódolása. A szövegek tehát nem egyszerűen csak kódolt betűk szekvenciái, hanem ezen felül struktúrával, tartalommal, valamint különböző olvasatokkal is rendelkeznek. A kódolás, illetve a jelölő nyelvek feladata e jellemzők explicitté tétele, hiszen a további megbízható – filológiai – feldolgozás csak ez úton válik lehetővé.

A „Repertorium” és a MASTER

Hiába az egységesítés, a szabványosítás magas-
tos célja, ha a *TEI* ajánlásainak eltérő értelme-
zéseiből fakadóan különböző elgondolások, s e
különböző elgondolásokon alapuló különböző
modellek születnek. A középkori kéziratok le-
írására szolgáló modellek egységesítésének cél-
jával jött létre a Bolgár Tudományos Akadémia
Irodalmi Intézetének „*Repertorium*” címet vi-
selő közös kezdeményezése. A projektben részt
vállalók közül két nevet mindenképp érdemes
megemlítenünk: *Anissava Miltenováét* és *David
J. Birnbaumét*. A kezdeményezés elsősorban a
középkori szláv kéziratok elektronikus leírására
koncentrál, de – ahogyan azt *Matthew Driscoll*,
a *TEI* egyik szakembere egy beszélgetés so-
rán e sorok írójának megjegyezte – a kéziratok
feldolgozásának problematikája voltaképpen
nyelvfüggetlen, avagy (származását, keletkezé-
sét nézve) bármilyen kézitről is legyen szó, a
felmerülő alapvető problémák azonosak.

Mielőtt a „*Repertorium*”-ról bővebb szó esne,
illő megemlítenünk a középkori kéziratle-
írások egyik alternatíváját, a *Manuscript Ac-
cess through Standards for Electronic Records*
(„Kéziratok Hozzáférhetőségének Biztosítása
Elektronikus Rekordokra vonatkozó Szabvá-
nyokkal”, továbbiakban: MASTER) elnevezé-
sű projektet, amelynek vezetői jeles középkor-
kutatókkal szoros együttműködésben, valamint
a *TEI* konzorcium szakemberei által alkotott
munkacsoporttal karöltve létrehoztak egy ja-
vaslatsomagot a kéziratok leírásának európai
szabványára vonatkozóan „*Towards a European
Standard for Manuscript Description*” címmel.
A szabvány kifejlesztésére szolgáló mag-DTD-
ként a tervezet a *MASTER DTD*-t javasolja, ame-
lyet speciálisan arra terveztek, hogy középkori
szövegek elektronikus leírásához szolgáljon al-
ternatív módszerekkel, amely módszerek – ál-
lítólagosan – az információkódolás lehetséges
megvalósításainak teljes tárházáról gondoskod-

nak. A kezdeményezés hátránya, hogy a szakér-
tők javaslataikat a nyugat-európai (angol, fran-
cia, német stb.) kéziratok leírására vonatkozóan
dolgozták ki.

Mind a *MASTER*, mind a *Repertorium* kezdemé-
nyezés a kéziratokra vonatkozó adatok strukturá-
lását a *TEI DTD*-k kiterjesztése, „újramodellezé-
se” révén valósítja meg. Mindez eltérő tervezési
stratégiákat és eltérő modelleket eredményez.
Az eltérés valószínűleg azzal magyarázható, hogy
a *Repertorium* nagyobb mértékben irányul egy
teljes – legalábbis részletesebb – kézirat-leírási
keret létrehozására, amelyből adatok nyerhetőek
ki egy rövid katalógus-leírás számára is, míglen
a *MASTER* eredendően egy rövid katalógus-le-
írás katalógus-adataira fókuszál inkább.

A két kísérlet tehát, akárcsak a legtöbb hasonló,
a kéziratok katalogizálásának szabványosításá-
ra irányul, nem a szövegstruktúrák tanulmányo-
zását lehetővé tevő teljes leírásokra. Ennek oka
egyfelől, hogy nem egyszerű feladat a közép-
kori kéziratok összes lehetséges szövegstruk-
túrájának leírására vonatkozóan konvenciókat
felállítani, másfelől e projektek a *TEI* ajánlása-
ira építve igyekeztek megalkotni modelljeiket,
ezen ajánlások azonban nagyobb mértékben tá-
mogatják a modern irodalmi szövegek, illetve
műfajok (vers, próza) leírását, mint a középkori
könyvtípusokét.

Fontos tehát kiemelnünk, hogy egyik modell
sem a kézratszövegek egészének kódolását tűzi
ki célul, s ilyesformán egyik sem járul hozzá a
bevezetőben említett korpusz létrehozásához.
Hogy filológiai jelentőségük mégiscsak meg-
kérdőjelezhetetlen, kiderül az alábbiakból.

A „Repertorium” története

A *Repertorium* voltaképpen egy a középkori
szláv kéziratok analitikus leírásaiból felépülő
információs adatbázis, amely részletes infor-
mációkkal szolgál a kéziratok paleográfiájáról,
kodikológiai, valamint textológiai adatairól, il-
letve tartalmazza az eredeti szövegek egy részle-

tét (a címet, az incipitet és az explicitet). A korpusz jelenleg a *Bolgár Tudományos Akadémia Irodalmi Intézetében* található.

1994-ben indult a „*Computer Supported Processing of Old Slavic Manuscripts*” („*Ószláv Kéziratok Számítógépes Támogatású Feldolgozása*”) címet viselő amerikai-bolgár projekt, amelyet többek között az *Open Society (Nyílt Társadalom)* alapítvány is támogatott. A nemzetközi munkacsoport a következő – David J. Birnbaum által megfogalmazott – öt alapelv követését tűzte ki célul:

1. a dokumentumok fájlformátumainak standardizálása;
2. többszöri felhasználás;
3. az elektronikus szövegek hordozhatósága (platformfüggetlenség);
4. a kéziratok elektronikus formában történő megőrzésének szükségessége;
5. az adatok jól-strukturált felosztása a szláv kodikológiai, ortográfiai, paleográfiai és textológiai hagyományok alapján.

A kezdeti szakasz eredménye egy a szláv kéziratok leírására alkalmas SGML-alapú modell létrehozása volt. (Az *SGML: Standard Generalized Markup Language*, vagyis egy szabványos jelölő nyelv, amelynek voltaképpeni leegyszerűsítése a WWW konzorcium későbbi fejlesztése: az *XML*.) A „*Template for Slavic Manuscripts*” egy DTD, amely a TEI SGML-gyakorlatának egyfajta kiterjesztéseként speciálisan a szláv kéziratok részletes leírását szolgálta.

1998-ig mintegy 200 középkori szláv kézirat feldolgozására került sor a Bolgár Tudományos Akadémia Irodalmi Intézetében. A projekt csatlakozott *Ralph Cleminson*nak a *Közép-Európai Egyetemen* létrehozott projektjéhez, amely a „*Computer-Supported Processing of Slavonic Manuscripts and Early Printed Books*” címet kapta (Cleminson 1998). Ez egyrészt a kéziratleírások számának növekedését, másrészt a technológiákra vonatkozó további cikkek megjelenését vonta maga után. A munka eredményeiről

az 1998-ban, Krakkóban megrendezésre került 12. Nemzetközi Szlavista Kongresszuson számoltak be. 2002 és 2003 között a kezdeményezés az „*Electronic Description and Edition of Slavic Sources*” („*Szláv Források Elektronikus Leírása és Kiadása*”) címet viseli, és elnyeri az UNESCO támogatását. Ebben az időszakban történik meg az áttérés az SGML-ről az XML-re. 2005-re 300-ra nőtt a leírt kéziratok száma.

Mit szolgál az XML-alapú leírás? (A leírás struktúrája)

Ha meg akarjuk válaszolni az alcímben feltett kérdést, sorra kell vennünk, mi az, ami a *Repertorium* projekt által javasolt mintán alapuló dokumentum-leírásból kiderül.

Maga a leírás hét fő összetevő elemből áll. Vegyük sorra ezeket!

1. A leírás legelején a katalogizálást végző személy gondoskodhat a kézirat formájára, típusára vonatkozó információkról (kódex?, tekercs?), valamint meghatározhatja, hogy fizikai és szellemi felépítését tekintve egy-egy mőről van-e szó, vagy sem. Ezután a *könyvtári információk* kerülnek kódolásra: itt adhatunk meg információkat arra vonatkozóan, hogy a kéziratot jelenleg hol őrzik (ország, város stb.) és melyik gyűjtemény részét képezi. Ugyanitt nyújthatunk információt arról, hogy a leírás alapjául más források szolgáltak-e.
2. Ezután következik az ún. *kodikológiai leírás*, amelynek első része tartalmazza a kézirat fizikai leírását; a főlíók számának meghatározásával kezdődik, majd a kézirat anyagára vonatkozó információ következik (ez általában pergamen vagy papír), ez után a lapok, illetve az írott részek méretei, az oszlopok és a sorok száma és a használt ábécé. Ezen adatok célja: rövid és lényegi információk

nyújtása a kéziratról. Egyes elemeket tovább bonthatunk: például az anyagra vonatkozó leíráson belül nyújthatunk egy általános jellegű leírást, majd meghatározhatjuk a pergamen vagy a papír típusát.

A *kodikológiai leírás második része* magában foglalja az összes kötésre, ívekre, fóliókra és oldalakra vonatkozó információt. Itt az összes ilyen jellegű jellemző megadható: például az ívek struktúrája, a hiányzó fóliók, a lapok elrendezése.

A *kodikológiai leírás harmadik része* a tintáról és a dekorációról tartalmaz információkat, különösen a dekoráció egyes elemeiről. Ez utóbbiakról további részletes leírások adhatók. A rongálódásra és a restaurációra vonatkozó információk is itt kerülhetnek feljegyzésre.

3. A másolat három fő jellemzőjének leírása képezi a feltárás következő, a *másoló* személyére vonatkozó részét, ezen jellemzők: a paleográfia, a nyelv és az ortográfia.
4. Ezután következik a *kézirat keletkezésének* helyére és idejére vonatkozó konklúziók levonása, amelyhez a megfelelő szintű kodikológiai, paleográfiai, ortográfiai és nyelvészeti megfigyelések nyújtanak alapot.
5. A *tartalom leírása* tekinthető a katalógus-információ magjának. A „*Repertorium*” modelljében ez az a hely, ahol a szövegre vonatkozó összes információ megtalálható, ezen belül: 1.) a kézirat egészének tartalmára vonatkozó információk; 2) a kézirat egyes szövegelemeire vonatkozó információk; 3.) a kéziratban található jegyzetekre vonatkozó információk.
6. A *történeti* részben a kézirat történetére, beszerzésére vonatkozó információkat találjuk.
7. az *adalek* információk között a reprodukcióra

(pl. fotózás, szkennelés), a kézirathoz csatolt, vele együtt őrzött anyagokra, a kódex kiállításokon való bemutatására vonatkozó feljegyzések találhatóak.

Mint láttuk, a *Repertorium* az ún. metaadatok leírását és kereshetőségének biztosítását tűzte ki célul, így a kézirat teljes szövege nem kerül feldolgozásra, mindössze az *incipit* és az *explicit*. A leírás nyelve az angol.

Ha visszakanyarodunk az eredeti problémához, vagyis a teljes szövegű korpusz létrehozásához a további filológiai kutatás érdekében, arra a következtetésre juthatunk, hogy ilyen korpuszra adott esetben talán azért nincsen szükség, mert a filológiai vizsgálódást éppen maguk a leírást folytató kutatók végzik, s a vonatkozó következtetések levonására már e tevékenység folyamán sor kerül. Nem véletlen, hogy a feldolgozás a Bolgár Tudományos Akadémia Irodalmi Intézetének égisze alatt zajlik, s olyan kutatók vesznek tevőlegesen részt benne, mint a projekt egyik vezetőjének számító, nagy szaktekintélynek örvendő szlavista professzornő: *Anissava Miltenova*.

Maga a munka amúgy hosszas és nehézkes, elég csak arra gondolnunk, hogy a kéziratok jelentős részét különböző (férfi-)kolostorokban őrzik, ami a feldolgozást kissé körülményessé teszi, hiszen a kéziratok helyben való leírására van csak lehetőség, ehhez pedig külön engedélyek szükségesek, s érthető kikötés (a pópák részéről), hogy nő ezt a munkát nem is végezheti. Meggyőződésem azonban, hogy a szláv nyelv- és irodalomtörténettel foglalkozó szakemberek részéről nem merül fel kétség az energia-befektetés hasznosságát illetően, hiszen a kéziratok (kódexek) feldolgozásának az egyszerű digitalizáláson (szkennelésen) túlmutató módja a metaadatokban való keresgélés, az információk összekapcsolása, „egymásra mutatása” révén rengeteg kiaknázható lehetőséget rejt magában.

Irodalom

1. BOJADŽIEV, A.: Electronic Student Editions of Medieval Slavic Texts. In: Scripta & e-Scripta, Vol. 1., Sofia, 2003.
2. BOJADŽIEV, A. – MILTENOVA, A. – RADOSLAVOVA, D.: A Unified Model for the Description of Medieval Manuscripts? In: Computational Approaches to the Study of Early and Modern Slavic Languages and Texts, Sofia, 2003.
3. BRADLEY, N.: Az XML-kézikönyv. Szak, Bicske, 2000.
4. BURNARD, L. Digital Texts, XML, and TEI - <http://www.tei-c.org/Talks/2005/Sofia/talk-intro.pdf>
5. CLEMINSON, R. Computer-Supported Processing of Slavonic Manuscripts and Early Printed Books. In: Annual of Medieval Studies in CEU, 1996–1997., Budapest, 146–148. p.
6. FLANDERS, J.: What the TEI does and how it works? Elhangzott: Azbuky Net Conference, Sofia, 2005. okt. 24-27.
7. KAGARLITSKY, Yu.: V. Internet Publication of Slavic Sources by the Russian Language Institute. In: Computational Approaches to the Study of Early and Modern Slavic Languages and Texts, Sofia, 2003.
8. Manuscript Access through Standards for Electronic Records - <http://www.cta.dmu.ac.uk/projects/master/>
9. MOTYGIN, O. V. – SLUTSKIJ, A. S.: Electronic Editions of Medieval Manuscripts: Thirteenth-Century Slavonic Euchologia. In: Computational Approaches to the Study of Early and Modern Slavic Languages and Texts, Sofia, 2003.
10. <oxygen/> XML Editor & XSLT Debugger - <http://www.oxygenxml.com/>
11. Repertorium - <http://clover.slavic.pitt.edu/~repertorium/>
12. SALGÁNÉ MEDVECZKI M.: Az XML. Új perspektívák a könyvtár-informatikában. In: Tudományos és műszaki tájékoztatás, 2004. (51. évf.), 2. sz., 61–71. p.
13. SINCLAIR, J.: Corpus, Concordance, Collocation. Oxford University Press, 1991
14. SÜTHEŐ P.: Hypertext. Természetes intelligencia az információtudományban. Budapest, Országos Széchényi Könyvtár, 1999.
15. Szeged Corpus: a natural language processed Hungarian corpus - <http://www.inf.u-szeged.hu/hlt>
16. [TEI] Text Encoding Initiative - <http://www.tei-c.org/>
17. [TEI Guidelines] Text Encoding Initiative Guidelines - <http://www.tei-c.org/P5/Guidelines/>
18. [TEI P4] Text Encoding Initiative Proposals 4 - <http://www.tei-c.org/P4X/>
19. Thesaurus Linguae Graecae - <http://www.tlg.uci.edu/>
20. VAKARELIYSKA, C. M.: Desiderata for an Electronic Collation of Medieval Slavic Gospel Texts. In: Scripta & e-Scripta, Vol. 1., Sofia, 2003
21. Extensible Markup Language - <http://www.w3.org/XML/>



Füzéki István Emlékdíj – A Magyar Könyvtárosok Egyesülete 2006. október 25-én konferenciát tartott az Országos Széchényi Könyvtárban „A mi 56-unk: könyvtár- és könyvtáros sorsok 1956-ban” címmel. Az ünnepi rendezvény alkalmából átadták a 2006. évi emlékérmét, melyet ez alkalommal Somkuti Gabriella kapott.