MIT JELENT A DIGITÁLIS FILOLÓGIA A SZEMANTIKUS WEB KORÁBAN? A DIGIPHIL PROJEKTRŐL

Palkó Gábor

PhD, tudományos titkár,
Petőfi Irodalmi Múzeum
palko.gabor@pim.hu

Ha meg akarjuk érteni, mit is jelent a digitális filológia a 21. században, történeti összefüggésbe kell helyezni a kérdést. Hogy miként változtak meg az elmúlt évtizedekben a hálózati kultúra (Manuel Castells) kialakulásával párhuzamosan a filológusok elvárásai, hozzáállásuk a számítógéphez, illetve hogyan alakult ki egy új interdiszciplína: a *computer humanities* és egy új kulturális jelenség, a *digitális kulturális örökség,* jelen keretek között nincs mód kifejteni (Palkó, 2015). Két csomópontot emelnék ki csupán ebből az izgalmas történetből. Hans Walter Gabler, amikor elkészült a híres és vitatott Ulysses-kiadással, kijelentette, hogy ezt a kiadást lehetetlen lett volna elkészíteni a számítógép segítsége nélkül. Ez még akkor is figyelemre méltó tény, ha tudjuk, a Gabler-féle nyomtatott kiadás valójában konzervatív elvekre épült, és távol áll a *nyitott szöveg* ideáljától, amelyet már a hetvenes években is összefüggésbe hoztak a számítógép filológiai szerepével (Gabler, 1989; McGann, 1989).

A másik, ha teszik, szimbolikus csomópont, amelyet kiemelnék a digitális filológia történetéből, voltaképpen még nem is nyert történeti távlatot. 2015-ben tette közzé Alois Pilcher, a bergeni Wittgenstein Archives (WAB) vezetője azt a Wittgenstein Nachlass-kiadást, amely már teljes egészében a szemantikus web technológiáján alapul (Tóth, 2010; Pichler, 2015). A WAB a kilencvenes évektől vezető szerepet játszott a digitális filológia világában a CD-kiadással és a központ akkori vezetője, Claus Huitfeldt publikációival (Huitfeldt, 1994/95).

Két éve indította útjára a Petőfi Irodalmi Múzeum – építve a Digitális Irodalmi Akadémia tapasztalataira – az MTA BTK Irodalomtudományi Intézetével együttműködve a DigiPhil projektet. Az itt szerzett tapasztalatok alapján – miközben a projekt jelenlegi és jövőbeli tartalmait, tevékenységtípusait és módszereit ismertetem –, megpróbálom kiemelni, hogy melyek a digitális kulturális örökség, azon belül a digitális filológia 21. századi gyakorlatának kulcsfogalmai, alapelvei és trendjei.

A DigiPhil-t – teljes nevén *Tudományos szövegkiadások, bibliográfiák és kutatási adatbázisok online tudástára* – az Irodalomtudományi Intézet szakmai partnerségével és a Grazi Egyetem *Zentrum für Informationsmodellierung* intézetével együttműködésben

alakítottuk ki. Jelenleg a DigiPhil (URL1) keretei között mintegy kétezer könyvoldalnyi tudományos szövegkiadás jelölőnyelvi átírását tesszük közzé, és ezer fakszimile oldalt szolgáltatunk nagy felbontású képek formájában. Az elkövetkező két-három évben a már rendelkezésre álló pályázati forrásokból mintegy tízezer fakszimile oldalt töltünk föl, illetve az ezekhez tartozó jelölőnyelvi szövegátírásokat készítjük el. A már feldolgozottak mellett további kritikai kiadásokat szeretnénk digitalizálni, egyik legfontosabb tervünk az Arany János emlékévre (2017) az Arany János kritikai kiadás digitalizálása.

A DigiPhil három párhuzamos alprojekttel kezdte meg a tudományos szövegek közzétételét. A Kassák Múzeum tudományos műhelye az avantgárd periodikák nemzetközileg is növekvő (el)ismertségű kutatóhelyévé vált az elmúlt években. A műhellyel együttműködve indítottuk el a Kassák Lajos által szerkesztett avantgárd folyóiratok *online* szolgáltatását. Jelenleg a szolgáltatás elérhetővé és kereshetővé teszi *A Tett* teljes anyagát fakszimilében és átírásban is. A következő időszak a tudományos jegyzetek elkészítéséé (a kutatócsoport annotált kiadást tervez). Az MTA–ELTE Kosztolányi kritikai kiadás műhellyel együttműködve tettük közzé az *Aranysárkány* kritikai kiadást, amely tartalmazza a regény kéziratának, kéziratos jegyzeteinek genetikus, a kézirat alkotásfolyamatát rekonstruáló komplex átírását, valamint az első nyomtatott kiadások betűszintű összeolvasását mintegy ezer oldal terjedelemben (Parádi Andrea, illetve Bengi László munkája). Hasonló terjedelmű és összetettségű az *Édes Anna* kritikai szövegkiadás, amely 2016 végéig kerül a nagyközönség elé online formában: a nyomtatott kiadás javított, bővített verziójaként. Megállapodtunk a kutatócsoporttal, hogy az összes általuk kiadott Kosztolányi kritikai kiadás a DigiPhil-en jelenik majd meg, az *Esti Kornél* kötet szintén az év végéig. Egy másik életmű kritikai kiadása, Mikszáth Kálmán összes művének két kötete, a 39-es és a 42-es szintén elérhető a DigiPhil weboldalán (URL1), és dolgozunk a 43. kötet közzétételén is. Ezek Mikszáth novelláinak különféle szövegváltozatait, azok összehasonlítását, illetve a keletkezésükre vonatkozó jegyzeteket, tanulmányokat tartalmazzák, Hajdu Péter rendezte őket sajtó alá.

A negyedik alprojekt, amely 2015-ben indult, szorosabban kötődik a Petőfi Irodalmi Múzeum gyűjteményéhez. Megkezdtük Móricz Zsigmond kiadatlan világháborús naplóinak *(Tükör)* feldolgozását, amely 4–5000 oldal terjedelmű kézírásos korpusz, s ennek annotált kiadását készítjük el. Különleges kihívás a textológusok számára, amikor nehezen olvasható szövegek átírását teszik közzé. A *Tükör* szöveg fakszimiléje önmagában nem teszi lehetővé, hogy a köz- és tudományos érdeklődésre számot tartó naplószöveg beépülhessen a Móricz-életmű olvasásába, hiszen a kézírás helyenként még az életmű szakértőjeként ismert és a projektet vezető Cséve Anna számára is nehezen olvasható. A jegyzetekkel ellátott átírás az egyetlen módja annak, hogy ez az értékes szöveg az irodalmi köztudat részévé válhasson. Hasonló célokkal ez év elején indul a Móricz Zsigmond-levelezés kritikai kiadása egy OTKA-projekt keretei között: mintegy tízezer levél adatainak publikálását és egy részük teljes szövegű közzétételét végezzük el. A fakszimile kiadás mellett betűhív átiratot, illetve kritikai jegyzeteket is készítünk.

Egy projektet mindig minősít, hogy saját módszertanát, álláspontját és rejtett elméleti előfeltevéseit képes-e reflexió tárgyává tenni.

Ennek tudatában – ez a DigiPhil ötödik alprojektje – vállalkoztunk arra, hogy a Ráció Kiadó engedélyével és a szerkesztők közreműködésével digitálisan is közreadjuk, és így a szélesebb közönség számára is elérhetővé tegyük a *Metafilológia* című hiánypótló fordításkötetet, amely mintegy 1500 oldalnyi filológiaelméleti szakirodalmat tartalmaz (Kelemen et al., 2014).

Az idáig felsoroltak képezik a DigiPhil szövegközreadó tevékenységének gyakorlati alapját, de a tudományos szövegkiadások, kritikai, illetve annotált szövegkorpuszok közreadása csak egy a projekt négy tevékenységi típusa közül. A második, amely a szövegek intelligens kereshetővé tételét szolgálja: az adatgazdagítás, illetve a kutatási adatok adatbázisba rendezése. Adatgazdagításon – a szemantikus web szemléletének megfelelően – a primer és szekunder szövegekben fellelhető nevek (személynevek, intézménynevek, földrajzi nevek) azonosítását és publikált névterekkel (tezauruszok, ontológiák) való összekötését értjük. Az irodalomtudományi relevanciájú adatokat, mindenekelőtt a bibliográfiai utalásokat adatbázisba rendezzük, ami olyan annotált (vagyis a szakértő által jegyzetekkel ellátott) bibliográfiák létrehozását teszi lehetővé, amelyek – a könyvtári adatkezelő szoftver terében létrejövő kapcsolati háló révén – egészen újszerű kérdésekre képesek választ adni. Arany János, József Attila és Kosztolányi Dezső bibliográfiájának készítését kezdjük el idén a kritikai kiadást készítő műhelyek szakértőinek segítségével. A harmadik tevékenységtípus az úgynevezett aggregáció. A DigiPhil projekt a világ legnagyobb kulturális tematikájú adatbázisának, az *Europeanának* aggregátora, vagyis az általunk szolgáltatott anyagokat leíró adatokat eljuttatjuk az *Europeana* szolgáltatásába, hogy minél szélesebb kutatói olvasóközösség érhesse el azokat. Ugyanakkor nemcsak a DigiPhil saját anyagairól gyűjtünk adatokat, hanem más, a publikációt saját eszközökkel végző szövegkiadó műhelyekkel is együttműködünk. Ennek az együttműködésnek a keretei között a nem a DigiPhil-en megjelent tudományos szövegkiadások metaadatait a saját oldalunkon közzétett anyagokkal együtt kereshetővé tesszük, illetve a metaadatokat eljuttatjuk különböző nemzetközi adatbázisokba, hogy ne csak a lokális közzététel helyén lehessen elérni őket, hanem szélesebb körben is kereshetővé váljanak. Jelenleg – a Klasszikus magyar irodalmi textológiai kutatócsoport aktív közreműködésével – Gyöngyössi János művei elektronikus forráskiadásának leíró adatait dolgozzuk fel. A tervünk az, hogy a minél szélesebb körből összegyűjtött adatokat (leíró, bibliográfiai adatok, teljes szövegek indexelt állományai, annotációk) közös felületen tegyük kereshetővé. A prototípus a DigiPhil keresőoldalán (URL2) érhető el.

A negyedik tevékenységtípus az annotáció. Természetesen a szövegkiadások maguk is tartalmaznak tudományos jegyzeteket, ám az ún. *szemantikus annotáció* eszközével a már publikált, zárt korpuszok utólagos, online és kollaboratív annotációja is lehetővé válik, méghozzá úgy, hogy ezek a megjegyzések a szövegekkel egyenértékű módon stabilan hivatkozhatóak, és intelligens kereséssel válnak elérhetővé. Történeti, régi szövegkiadások vagy akár periodikák válhatnak úgy jegyzetelhetővé, hogy a munkát a kutatócsoportok szakértői online eszközökkel távolról végzik.

***

A következőkben három olyan problémakört érintek, amelyek – bár nem jelentenek újdonságot a számítástechnika történetében –,

a digitális kulturális örökség 21. századi világában új megvilágításba kerülnek. Az első ilyen probléma az adatbiztonság és a hosszú távú megőrzés kérdése. Az anyagi hordozó megsemmisülésének veszélye mindig kísértette az írásrendszerekre épülő kultúrákat, ám az elektronikus jeltovábbítás korában a probléma egészen új arcát mutatja. Az adatok, mivel a számítógép fekete doboza nem enged közvetlen bepillantást a tárolás konkrét lokalitásába, láthatatlanul íródnak és törlődnek. A számítástechnika hőskorában ezért az adatvesztéstől való félelem uralta a digitális kulturális örökség világát. A virtuálisan korlátlan tárolókapacitás, a felhőalapú számítástechnika jelenében sokkal inkább az észrevétlen adatmódosulástól, a véletlen újraíródástól kell tartanunk. Mindez arra a köznapi tapasztalatra emlékeztet, amikor különböző számítógépeken írjuk „ugyanazt" a dokumentumot, és egy adott pillanatban képtelenek vagyunk eldönteni, hogy melyik az aktuális verzió a kópiák sokaságából.

A gyorsuló ütemben növekvő digitálisan hozzáférhető dokumentummennyiség egy másik veszélyt is magában rejt: hogyan biztosítható, hogy azok a szövegek, amelyeket közzéteszünk, megtalálhatóak legyenek a verbális információ tengerében. Mindkét problémára léteznek kész megoldási módszerek. A szemantikus technológiák az utóbbi veszélyre (állapotra?) adott válaszként is értelmezhetőek: a szabványos és hierarchikusan strukturált, hálózatba rendezett metaadatokkal radikálisan javítható a releváns találatok aránya. Az adatvesztés és észrevétlen adatmódosulás veszélyétől pedig az aprólékosan kidolgozott ajánlások óvhatnak bennünket, amelyeket jelentős memóriaintézmények, illetve intézményi konzorciumok alkottak meg a digitális adatok védelmére. A DigiPhil projekt a *Data Seal of Approval Guidlines* (DSA) ajánlásainak megfelelően építette ki belső adatbiztonsági szabályzatát. Ez elsősorban az infrastruktúra felépítésére vonatkozik, az ezen belüli egyes feladatokat célszoftverek végzik. A munka kezdetétől verziókövető szoftver segítségét vesszük igénybe, hogy ne csak az aktuális verziót lehessen mindig megtalálni, hanem vissza is lehessen követni a szövegek módosulásait. A módosítások mindegyike információt tartalmaz a rendszerben a módosítást végrehajtó személyről és a változtatás időpontjáról is. A közzétételre kész verziót ezután repozitori (repository) szoftverbe töltjük át, amely kifejezetten digitális objektumok kezelésére használatos célszoftver: az objektumok adminisztrálására, archiválására, karbantartására, metaadatolására, közzétételére és kereshetővé tételére terveztek. A repozitori szoftverek maguk is képesek verziókövetésre és archiválásra is, de biztonságosabb ezen funkciók szétválasztása. Magát az archiválást – az adatbiztonsági ajánlásoknak megfelelően – több lépcsőben kell végezni, napi, heti, havi ciklusokat különböztetve meg. Az adatokat adott periódusonként hosszú távú adatmegőrzésre kialakított, nem hálózati eszközökkel is érdemes archiválni: jelenleg a szalagos meghajtó az egyik legjobb megoldás erre a célra, melyet a DigiPhil is használ.

A 21. században a digitális kulturális örökség világában talán legerősebb elvárásként a platformfüggetlenség, nyílt hozzáférés (Open Access) és szabványosság kívánalmai fogalmazódnak meg. (Az Európai Uniós pályázatok gyakran a támogatás feltételeül szabják, hogy nyílt forráskódú, ingyenes és szabványos eszközöket kell használni, illetve fejleszteni.) A DigiPhil is ezt az utat követi: a szolgáltatás építőkockáit jelentő célszoftverek mindegyi-

ke nyílt forráskódú, ingyenes; kulturális intézmények, egyetemek nemzetközi kutatói közösségei által kifejlesztett és karbantartott eszköz. A digitális filológia hajnalán Gabler még egyetlen komplex szoftver segítségével gondolta megvalósíthatni az összes textológiai funkciót – mára ez a modell idejétmúlttá vált. A tucatnyi együttműködő kisebb programból felépített szolgáltatás azonban csak akkor fenntartható, ha a komponensek együttműködése szabványos felületeken, szabványos feltételekkel folyik, így az egyes komponensek adatvesztés nélkül kiválthatóak.

A digitális filológiában évtizedek alatt egyeduralkodóvá vált az XML dokumentumleíró szabvány és az arra épülő Text Encoding Initiative (TEI) ajánlás, amelyet kifejezetten bölcsészettudományi használatra, textológus szakemberek közössége hozott létre (Renear, 2004; Bíró, 2005; Kalcsó, 2011). A kilencvenes években még komoly viták folytak arról, hogy milyen hátrányai lehetnek a dokumentumok leírására kifejlesztett hierarchikus jelölőnyelveknek, de a TEI végül alternatíva nélkül maradt (Renear et al., 1996). A DigiPhil projekt kiépítése során kiderült, hogy önmagában a jelölőnyelvi szabvány használata nem elégséges, mert a szövegtest és a szövegjelenségek kódolásán (erre való a TEI XML) a szövegekre vonatkozó adatok, a metaadatok szabványosítása is elengedhetetlen. Erre vonatkozóan a TEI ajánlása nem elég szigorú és szisztematikus. A digitális filológiai műhelyek különböző megoldásokat hssználnak a metaadatok szabványosítására. A DigiPhil – követve a Wittgenstein Source példáját–, a Digital Manuscripts to Europeana (DM2E) projekt által kidolgozott metaadat-kezelési szabványt és munkafolyamatot alkalmazza. A fenti projekt résztvevőjeként a Kassák-szerkesztette *A Tett* szövegére vonatkozó metaadatokat eljuttattuk az Europeanába – Magyarországon először gyakorolva az aggregációt filológiai adatokon.

Az aggregáció a digitális kulturális örökség legmeghatározóbb trendje, melynek során a kisebb helyi gyűjtemények digitális anyagait, képeket, hangokat, videókat és szövegeket, illetve az ezeket leíró adatokat nagyobb, nemzeti vagy nemzetközi szolgáltatásokban összegyűjtik, feldolgozzák, gazdagítják, kereshetővé teszik. A legnagyobb volumenű és hatású aggregátor a világon az Europeana, melyben jelenleg 49 millió digitális objektum található. Egy ekkora adatbázis kezelése, az abban való eligazodás nem könnyű, sokan vitatják is az ilyen típusú globális projektek használhatóságát. Ugyanakkor az a metaadatkezelési metódus, amelyet az Europeana-hoz kötődő pályázatok (például: Athena, Linked Heritage, DM2E) kialakítottak, illetve maga az Europeana által kifejlesztett EDM-szabvány, úgy tűnik, a digitális kulturális örökség minden területén teret nyer. Ennek a szabványglobalizációnak a legjobb példája, hogy az Egyesült Államokban is létrejött az Europeana adatmodelljének mintájára egy projekt, a Digital Public Library of America (DPLA), amelyben pillanatnyilag 11 millió objektum található.

A digitális filológiai adatok aggregációja speciális terület, az Europeana anyagának döntő többsége ugyanis kép, vagyis digitális képpel jól reprezentálható objektum (képzőművészeti alkotás, fotó stb.). A DM2E projekt célkitűzése az volt, hogy kidolgozza az Europeana metaadatsémájának egy, a filológiai objektumokra szabott verzióját, és erre építve aggregációs munkafolyamatot és intézményi közösséget építsen ki. Ez annál is fontosabb, hiszen a digitális filológia világában jelenleg aktív aggregátorok többsége nemzeti szintű adatgyűjtést végez: például német nyelvterületen a TextGrid és a Deutches Textarchive (DTA), nemzetközi aggregációt végez a kisebb volumenű TAPAS projekt, valamint a University of Oxford Text Archive. Ezen projektek mind TEI XML-ben kódolt szövegeket gyűjtenek össze. A Deutches Text Archive a maga 130 000 digitális objektumához saját, rendkívül kifinomult, egyedi metaadat-szintaxist dolgozott ki, amely ugyanakkor nem támogatja az aggregációs folyamatokat. A DigiPhil – mint szövegeket közzétevő és aggregáló szolgáltatás – a DM2E modelljét követi, amely már eleve különféle forrású adatok összehangolására jött létre.

A különféle forrásokból származó adatok összegyűjtésének igazi tétje, hogy sikerül-e a lokális gyűjtemény kontextusát átlépve relevánsabb és intelligensebb keresési metódusokkal új kutatói érdeklődésnek, új témáknak vagy módszereknek megfelelő találati halmazokat létrehozni. A DigiPhil projektben a teljes szövegű keresés és az úgynevezett facettált böngészés *(faceted browsing)* technológia házasítását alkalmazzuk erre a célra. A facettált böngésző segítségével előre adott szempontok alapján szűkíthetjük az (aggregált adatoknál többnyire kezelhetetlenül nagy) találati halmazt. A szűkítés szempontja az objektumokat leíró bármely adattípus lehet: a szövegek szerzője, a levelek címzettje, a keletkezés helye vagy dátuma stb. Egy kutatási szempontból releváns kérdés (például hogy egy adott évben az adott szerző hol publikált) néhány kattintással eredményt ad. Ezt a metódust kombináljuk a teljes szövegben történő kereséssel: szöveg, szó vagy szótöredék megadásával, illetve csonkolás és a joker karakterek használatával is. Hamarosan a szinonimákra és a szótő alapján történő keresést is integráljuk. A jelölőnyelvi kódolás ennél speciálisabb kereséseket is lehetővé tesz. Mivel a TEI XML jelölőnyelvében a filológusok által fontosnak tartott minden textuális jellemző, illetve szövegművelet jelölhető (a szerzői törléstől és beszúrástól a szövegromláson át az utólagos szerkesztői kiegészítésekig stb.), rendkívül hasznos – ezt még a nagy digitális filológiai projektek közül is csak néhány teszi lehetővé –, ha az egyes szövegjellemzőkre is lehet keresni. Érdekes kutatói kérdés lehet például, hogy egy adott szerző vagy irodalomtörténeti mozgalom kéziratain belül melyek voltak a legtöbbször törölt/javított szavak.

Ez még csak a kezdet. Az egyre növekvő korpusz, elsősorban a teljes életművek kritikai feldolgozása lehetővé teszi majd stilisztikai, szóstatisztikai vizsgálatok elvégzését is. Kidolgozott módszertanok állnak rendelkezésre ilyen elemzésekhez, amelyek alkalmazására eddig Magyarországon csak szűk irodalmi korpuszokon történt kísérlet. A projekt továbbfejlesztésének következő iránya az adatvizualizáció, amely jelentheti például a szógyakorisági statisztikákat vagy földrajzi adatok, keletkezési helyek és időpontok megjelenítését korabeli térképeken, de intertextuális kapcsolatok hálózatának megjelenítése is lehetségessé válik – amelyre a Wittgenstein Source egy projektje ad kitűnő példát. Terveink szerint automatikus intertextus-keresést is végzünk a közeli jövőben a feldolgozott szövegekben: a plágiumkereső technológia ugyanis megfelelő paraméterezés mellett az irodalmi szövegek idézettechnikájáról is értékes információkat nyújthat.

Kulcsszavak: *kritikai kiadás, textológia, szemantikus web*

IRODALOM

Bíbor Máté et al. (2005): *A magyar irodalom filológiája*. Gépeskönyv • http://www.tankonyvtar.hu/en/tartalom/tkt/magyar-irodalom/index.html

Bíró Szabolcs (2005): *Szövegfeldolgozás XML alapokon*. Neumann Kht., Budapest • http://www.tankonyvtar.hu/informatika/szovegfeldolgozas-xml-080906-159

Gabler, Hans Walter (1989): A kiadói szöveg születése: számítógép bába-szerepben. (ford. Farkas Ildikó) *Helikon.* 3–4, 421–428.

Huitfeldt, Claus (1994/95): Multi-Dimensional Texts in a One-Dimensional Medium. *Computers and the Humanities.* 28, 235–41. DOI: 10.1007/BF01830270

Kalcsó Gyula (2011): *A TEI-XML felhasználása magyar nyelvű korpuszok építésében.* In: Boda István Károly – Mónos Katalin (szerk.): *Az alkalmazott nyelvészet ma: innováció, technológia, tradíció. XX. Magyar Alkalmazott Nyelvészeti Kongresszus.* Debrecen, Manye–Debreceni Egyetem, Budapest. 65–71. • http://www.inf.unideb.hu/~bodai/pub/MANYEXX_elso_resz-B5.pdf

Kelemen Pál – Kulcsár Szabó E. – Tamás Á. – Vaderna G. (szerk.) (2014): *Metafilológia 2. Szerző – könyv – jelenetek*. Ráció, Budapest

McGann, Jerome J. (1989): Az *Ulysses* mint posztmodern szöveg: a Gabler-féle kiadás. (ford. Friedrich Judit) *Helikon.* 3–4, 429–452.

Oliver, Andrew (1989): Mikroinformatika és textológia. (ford. Farkas Ildikó) *Helikon.* 3–4, 412–420.

Palkó Gábor (2015): Digitális filológia: számítógép anyaszerepben. *Filológiai Közlöny.* 2, 187–199. • http://www.balassikiado.hu/BB/NET/Filologia/Filopdfek/Filo_2015_2.pdf

Pichler, Alois (ed.) (2015): *Wittgenstein Source Bergen Nachlass Edition*. Edited by the Wittgenstein Archives at the University of Bergen under the direction of Alois Pichler. In: Wittgenstein Source (2009–). WAB, Bergen • www.wittgensteinsource.org

Renear, Allen H. (2004): Text Encoding. In: Schreibman, Susan et al. (eds.): *A Companion to Digital Humanities*. Blackwell, Oxford • http://www.digitalhumanities.org/companion/

Renear, Allen – Mylonas, E. – Durand, D. (1996): Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies. In: Ide, Nancy – Hockey, Susan (eds.): *Research in Humanities Computing*, Oxford University Press, Oxford. 263–280. • http://cds.library.brown.edu/resources/stg/monographs/ohco.html

Tóth Máté (2010): Könyvtárak a szemantikus web világában. *Könyvtári Közlöny*. 3, 413–438. • http://ki.oszk.hu/kf/2010/10/konyvtarak-a-szemantikus-web-vilagaban/

URL1: digiphil.hu

URL2: data.digiphil.hu/search/

A BIG DATA KIHÍVÁS ÉS LEHETŐSÉG A BÖLCSÉSZETTUDOMÁNYOKBAN: DIGITÁLIS SZÖVEGEK ÉS METAADATOK TÁVOLI OLVASÁSA

Péter Róbert

egyetemi adjunktus,
Szegedi Tudományegyetem Angol Tanszék
rpeter@lit.u-szeged.hu

Mindannyian tudjuk és tapasztaljuk: az információs társadalom korszaka jelentős kihívás a humán tudományok számára: a változás egyaránt érinti a tudás társadalmi intézményeit, legitimációját, ám ugyanúgy kihat mindennapjainkra, tudományos gyakorlatunkra. Mindez alkalmat nyújt – valójában sürget – a humántudományi kutatások tárgyának, módszerének és közegének újragondolására.[1]

A bölcsészettudományokban zajló digitális fordulatot jelzi, hogy az elmúlt évtizedben hatalmas mennyiségű ismert és ismeretlen forrásanyag vált elérhetővé és kereshetővé szabad felhasználású vagy előfizetést igénylő digitális gyűjteményekben. Nyilvánvaló, hogy a digitális forradalom jóval gyorsabbá, könnyebbé tette a kutatást azzal, hogy több millió szöveget tartalmazó digitális archívumokban kereshetünk, korábban nem, vagy nehezen hozzáférhető anyagok váltak elérhetővé. Az angol nyelvű írott kultúrkincsre vonatkozóan Matthew L. Jockers 2008-at, a németre vonatkozóan Fotis Jannidis és Gerhard Lauer pedig 2011-et jelöli meg az áttörés éveként (Jockers, 2013; Janidis – Lauer, 2014). Több millió szöveg azonban nemcsak könnyebbséget jelent: a digitális fordulat valójában új kihívások elé állította a kutatókat. A gyakorlatban végbement digitális forradalmat azonban még nem igazán követte módszertani forradalom. A lehetőségeket és kihívásokat is rejtő digitális kutatás jelenlegi helyzetét tökéletesen példázza a verseny, melyet a British Library hirdet meg 2013 óta: a British Library Labs pályázatot ír ki tudósoknak, kísérletező szakembereknek és szoftverfejlesztőknek, melynek célja a neves brit intézmény digitális gyűjteményeinek felhasználására épülő innovatív és úttörő projektek megvalósítása (URL2). Olyan újszerű kutatási elképzeléseket várnak, melyek hiánypótló ismeretek, felfedezések forrásaként támaszkodnak a könyvtár hatalmas digitális gyűjteményeire. A digitális tudományos kutatás valóban formabontó és úttörő kutatási lehetőségekkel kecsegtet a bölcsészettudomá-

[1] Kokas Károly, Labádi Gergely, Péter Róbert, *Digitális bölcsészet Szegeden* – konferenciafelhívás (2015. október 12) URL1. Ezúton mondok köszönetet Labádi Gergely kollégámnak e tanulmány korábbi verziójához fűzött építő és értékes megjegyzéseiért, valamint azért, hogy utalhattam *A magyar regény adatbázisa* című projekt eddig publikálatlan eredményeire.