

A DIGITÁLIS KRITIKAI SZÖVEGKIADÁSOK KIHÍVÁSAI ÉS LEHETŐSÉGEI

A digitális környezet hatása a filológia hagyományos folyamataira

Mihály Eszter

mihaly.eszter@oszk.hu

Magyar Nemzeti Múzeum Közgyűjteményi Központ Országos Széchényi Könyvtár,
Digitális Bölcsészeti Központ, Digitális Filológiai és Webarchiválási Osztály, osztályvezető

Absztrakt

Hogyan zajlik egy digitális kritikai szövegkiadás készítése? Milyen pontokon kell kapcsolódnia a kutatócsoport és a digitális bölcsészek munkájának? Mennyiben tér el ez a munkafolyamat egy hagyományos, papír alapú kritikai kiadás készítéséhez szükséges lépésektől? Milyen eszközöket és szabványokat szükséges használni a munkához? Mi az Entitástár és mi a létjogosultsága? Mennyiben más a szövegkiadás módszertana, ha írói levelezéskiadásról van szó? Az előadás ezeket a kérdéseket járja körül az MNMKK OSZK és az MNMKK PIM Kassák Múzeum együttműködésében készülő Kassák Lajos levelezésének online kritikai kiadása tapasztalatai alapján.

Kulcsszavak: digitális bölcsészet, digitális filológia, tudományos szövegkiadás, entitás, írói levelezések

Abstract

Challenges and Opportunities of Digital Scholarly Editions: The Impact of the Digital Environment on Traditional Philological Processes

How is a digital scholarly edition created? At which points should the work of the research group and digital humanities experts be connected? How does this workflow differ from the steps required for a traditional, print-based critical edition? What tools and standards are necessary for the work? What is the Entity Repository, and what is its justification? How does the methodology of text editing differ when it involves the publication of an author's correspondence? This paper explores these questions through the experience gained during the preparation of the online edition of Lajos Kassák's correspondence, created in collaboration between the MNMKK OSZK and the MNMKK PIM Kassák Museum.

Keywords: digital humanities, digital philology, scholarly edition, entity, writers' correspondence

Bevezetés

A filológia hagyományos módszertana évszázadokon keresztül a szövegek manuális feldolgozásán, értelmezésén és kiadásán alapult. Az utóbbi évtizedek digitális fordulata azonban nem csupán eszköz-, hanem szemléletváltást is eredményezett. A kritikai szövegkiadás új formája, a digitális kritikai- vagy tudományos kiadás egyszerre épít a filológia hagyományaira és a digitális infrastruktúrák rugalmasságára.

1. A digitális kritikai/tudományos kiadás keretrendszere

A digitális kritikai szövegkiadás olyan összetett tudományos tevékenység, amely nem csupán digitalizál, hanem strukturál, összekapcsol és értelmez is. Ennek megfelelően a digitális kiadások több szinten is jelentősen különböznek a nyomtatott változatoktól: a felhasználói hozzáférés, a kereshetőség, az interaktivitás és a szemantikus adatkapcsolatok új lehetőségeket teremtenek az értelmezésre és az adatfeldolgozásra.

A digitális kiadás nem kizárólag az olvasói szokásokhoz és technológiai változásokhoz igazodó publikációs forma, hanem kutatómódszertani fordulatot is jelent, hiszen magában foglalja az adatmodellezés, az adatbázis-építés, a vizualizáció, a szemantikus web és egyéb technológiák integrációját. A digitális kritikai kiadásokban tehát nemcsak a szöveg, hanem a köré épülő információs hálózat is központi szerepet kap, s ennek köszönhetően jóval nagyobb mértékű *adatközpontúság*¹ jellemzi.

Egy digitális kiadási projekt elindítása emellett interdiszciplináris együttműködést is feltételez: a kutatók, filológusok, informatikusok, digitális bölcsészek közösen dolgoznak egy egységes rendszer létrehozásán. Már a kezdeti fázisban szükség van egy a bölcsészettudományok és az informatika közötti szakadékot átívelő közös nyelv kialakítására, amely lehetővé teszi az elméleti és technikai szempontok harmonizálását. E nyelv nemcsak a terminológia egységesítését, hanem a technológiai és módszertani összehangolást is szolgálja.

A dHUpla (Digital Humanities Platform)² környezetében publikálandó digitális filológiai projekt elindításához ennek megfelelően a következő alapvető lépésekre van szükség:

- Célok és mérföldkövek kijelölése, ütemterv és projektterv elkészítése;
- Textológiai-filológiai alapelvek kidolgozása;
- XML-tagkészlet meghatározása (dHUpla-specifikáció alapján);
- Szükséges fejlesztések meghatározása;
- Útmutatók készítése, eszközök, módszerek betanítása;

1 vö. Maróthy Szilvia, *A vizuális és az adatközpontú szemlélet* in: Helikon, 2020/1. 5–15.
https://helikonfolyoirat.hu/wp-content/uploads/2021/12/Helikon_2020-1-5-15.pdf

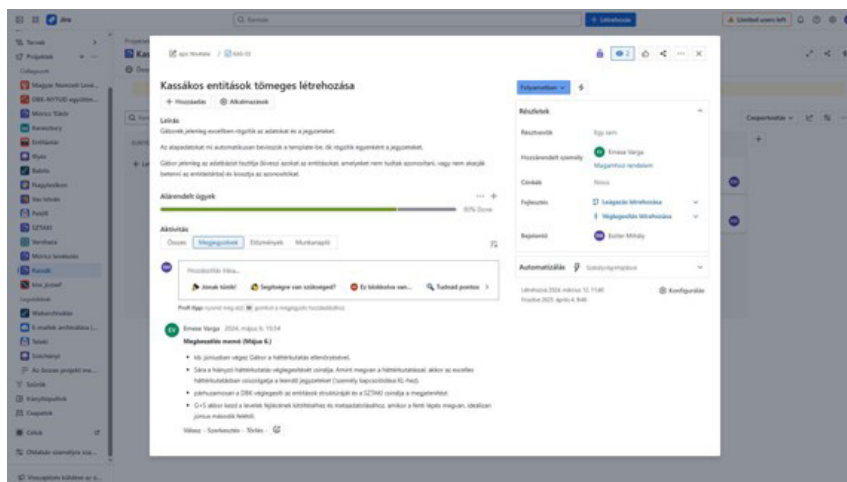
2 <https://dhupla.hu/> Hozzáférés: 2025. 08. 07.

- Feldolgozandó anyag megismerése, tanulás;
- Kommunikációs csatornák kiépítése.

A projektmenedzsmentben két kulcsfontosságú eszközt alkalmazunk. A Jira³ segítségével a résztvevők nyomon követhetik a feladatok előrehaladását, határidőket állíthatnak be, valamint az agilis módszertan alapján sprintalapú munkavégzést is megvalósíthatnak; a Trello-t⁴ pedig ún. content-management eszközként használjuk az előállított tartalmak nyilvántartásához és kezeléséhez.

A használt eszközök funkciói tehát:

- **Jira:** feladatok, határidők, státuszok kezelése, agilis módszertan alapján;
- **Trello:** vizuális tartalommenedzsment és kollaboráció, projektállapotok átlátható követésére.



Ez az integrált, technológiailag támogatott együttműködés nem csupán hatékonyságot eredményez, hanem minőségi garanciát is jelent a kutatási projektek szempontjából. A transzparens munkamegosztás, a közös döntéshozatal és a jól dokumentált szerkesztési folyamatok révén a digitális kiadás megbízható és hosszú távon fenntartható tudományos produktummá válik.

A rendszer ezenkívül lehetővé teszi, hogy a különböző szakterületek képviselői valós időben kövessék a projekt előrehaladását, visszacsatoljanak, és közösen formálják az eredményeket.

3 <https://www.atlassian.com/software/jira> Hozzáférés: 2025. 08. 07.

4 <https://trello.com/> Hozzáférés: 2025. 08. 07.

2. Szöveg-előállítás MI-alapokon: HTR-modellek

A digitális szövegkiadás egyik technológiai sarokköve a gépi kézírás-felismertetés, amely a történeti dokumentumok digitalizálását és szöveges feldolgozását támogatja. A HTR (Handwritten Text Recognition) technológia segítségével lehetőség nyílik a kéziratok automatizált felismertetésére és karakterpontos átírására. Az így előállított szövegek alapjául szolgálnak a további nyelvészeti, történeti vagy irodalomtudományi elemzéseknek.

A mesterségesintelligencia-alapú HTR-modell képes megtanulni az egyes írásképekhez tartozó karaktereket, szavakat, sorokat, mintázatokat, s a szövegkiadási projektek előállt szövegkorpuszai folyamatosan hozzáadhatóak a tanítókorpuszhoz, amely összesítve egyre nagyobb hatékonyságú modellt tud előállítani. A digitális bölcsész egyik legfőbb felelőssége összefogni a különböző filológiai projekteket, hogy *az adatok és a tudás ezen a területen is összeadódjon (tudás- és adatmegosztás 1.)*

Az általunk fejlesztett legújabb „Magyar vegyes 2.0”⁵ modell vegyes kézírású adatkészleten tanult (Kiss József, Móricz Zsigmond, Egry József, Kassák Lajos, Simon Jolán kézírása), amely mindenki számára elérhető a Transkribus felületén. A modell 798 dokumentumon, összesen 3393 oldalon és 322 542 szón végzett betanítás után nagy előrelépést mutatott: a karakterhiba-arány (CER) 6,16%-ra csökkent, ami jelentős minőségi ugrást jelent a korábbi, 10,7%-os értékhez képest. Ez a fejlődés lehetővé teszi a nagy mennyiségű magyar nyelvű, 19–20. századi kéziratos forrásanyag gyors és pontos feldolgozását, ami kulcsfontosságú a digitális kiadások időhatékony létrehozásához.

3. Strukturált szövegkódolás: a dHUpLa-tagkészlet

A digitális szövegkiadások sikerességének másik kulcsfontosságú tényezője a megfelelő és egységes szövegkódolási rendszer kiválasztása és adaptálása. A magyar nyelvi sajátosságokra épülő dHUpLa-tagkészlet specifikációja⁶ egy olyan TEI-kompatibilis⁷ ajánlás, amely alkalmas a magyar szövegek szerkezeti, nyelvi és filológiai jellemzőinek egységes és strukturált reprezentálására. A részletes, ám rugalmas szabályrendszer célja nem csupán a szövegek gépi olvashatóságának biztosítása, hanem a tudományos feldolgozhatóság elősegítése is.

A „Tollvonásból kódrendszer” szlogenünk szimbolikus jelentéssel bír: arra utal, hogy a történeti kéziratokban megjelenő szövegeket akár az utolsó tollvonásig digitálisan strukturált, formális szabályokon alapuló XML-reprezentációvá alakítjuk. A TEI XML-kódolás lehetővé teszi az előző pontban előállt digitális szövegben a tartalom és a szerkezet elvá-

5 <https://app.transkribus.org/models/public/text/hungarian-handwriting-19th20th-century-20>
Hozzáférés: 2025. 08. 07

6 <https://dhupla.hu/page/muhely/dhupla-spec> Hozzáférés: 2025. 08. 07.

7 <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html> Hozzáférés: 2025. 08. 07.

lasztását, különböző textológiai–filológiai annotációk beágyazását, s a projekt által megkívánt mélységű, illetve jellegű adatgazdagítást. Ehhez a szerkesztői munkához fejlesztjük a SZTAKI-val együttműködésben, a folyó projektek igényeire folyamatosan reagálva a dHUpa-framework-öt⁸, a dHUpa felhasználóbarát szerkesztői környezetét.

4. Entításalapú szemlélet és Entitástár

A digitális filológiai munka adatkezelésének kulcseleme az entításalapú adatmodellezés. A dHUpán létrehozott Entitástár⁹ olyan filológiai szemléletű szemantikus adattár, amely lehetővé teszi, hogy a szövegekhez kapcsolódó fogalmi elemek (pl. személyek, helyszínek, események, művek) strukturált módon és kifejezetten a szövegkiadásokhoz kapcsolódó információkkal ellátva legyenek nyilvántartva és összekapcsolva. Az entitásokat nem csupán leírásokkal, hanem állandó egyedi azonosítóval is ellátjuk, amely alapján egyértelműen beazonosíthatók különféle digitális környezetekben. A dHUpa Entitástára tehát projekt-specifikus adatok és leírások gyűjtőhelye, nem általános, hanem filológiai aspektusú szaknévtér.

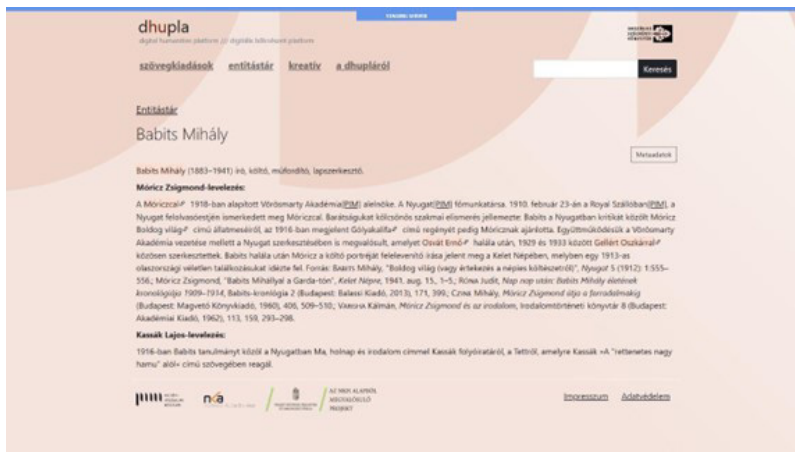
Alapját szintén a TEI XML-alapú struktúra képezi, amely rugalmasan képes kezelni különféle entitástípusokat. Minden entitás egy önálló XML-fájlban kap helyet, amely tartalmazza annak nevét, típusát, projektenként elválasztott leírását, valamint a kapcsolódó belső és külső azonosítókat (pl. PIM Névtér, Wikidata, VIAF, Geonames). A tartalom XPath-tal könnyen lekérdezhető, így nem szükséges hozzá külön adatbázis, s teljesen platformfüggetlen. Ezenkívül az XML-formátum a legalkalmasabb strukturált adatok és szöveges tartalom együttes tárolására, továbbá az ember és gép számára is könnyen feldolgozható. Az entitások szerkesztését a szövegekhez hasonlóan a dHUpa-frameworkben végezhetik a munkatársak, a verziókövetés és közös szerkesztés Gitlab-on keresztül történik.



8 Mihály Eszter és Micsik András, *Szerkesztői környezet TEI-alapú szövegkiadásokhoz*. In: Új technológiákkal, új tartalmakkal a jövő digitális transzformációja felé. HUNGARNET Egyesület, 2023, Budapest, 186–191. ISBN 9786158224314 <https://doi.org/10.31915/NWS.2023.27>

9 <https://dhupla.hu/collection/entitastar> Hozzáférés: 2025. 08. 07.

Az ún. staging szervert lehetővé teszi az éles publikálás előtti ellenőrzést, így a tartalmak minősége és konzisztenciája biztosítható.



A szemantikus modellalkotás legnagyobb előnye *az adatok újrafelhasználhatóságában és összekapcsolhatóságában rejlik (tudás- és adatmegosztás 2.)*. Egy személy, aki több különböző projekthez is kapcsolódik, csak egyszer szerepel az Entitástárban, elkerülve a redundanciát, valamint összegyűjtve a különböző kutatásokból származó eredményeket. A kutatók így kooperálhatnak is a hiteles információk előállításában, s az entitások közötti kapcsolódások révén nemcsak az egyes entitásokat, hanem a közöttük lévő relációkat is vizsgálhatják.

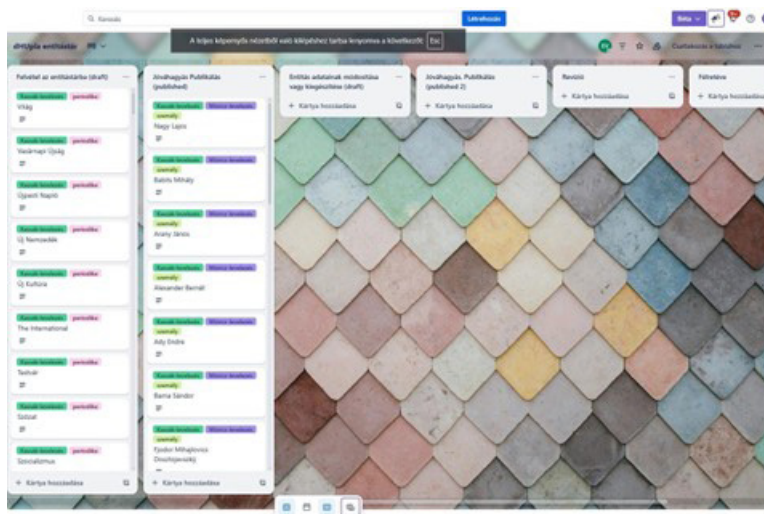
Miért kiemelten fontos tehát a digitális szövegkiadásban az entitás mint adat? A válasz a multifunkcionalitásában rejlik:

- Beazonosít, egyértelműsít;
- Általános és specifikus jellemzők egyidejű, többszempontú halmaza;
- Projekteken átível, összeköt;
- Megszünteti a párhuzamosságokat;
- Keresések, lekérdezések, statisztikák forrása;
- Szemantikus kapcsolatok, vizualizációk alapja.

Az adatok megszületéséhez a tudományos szövegkiadások során a textológiai–filológiai munka háttérében minden esetben kutatómunka áll. A digitális kiadás esetén ennek a háttérkutatásnak (pl. személynevek azonosítása) már a kezdetektől a digitális filológiai szempontokkal, elvárásokkal összeegyeztetve kell szerveződnie. Az adott kutató/kutatócsoport természetesen végezhet folyamatos háttérkutatást saját, megszokott eszközökkel és módszerekkel (pl. papíron, táblázatokban stb.) a következőképpen:

- Előzetesen egyeztetett, géppel feldolgozható formában, struktúrában kell készülnie;
- A megfelelő időpontban (meg egyezés szerint¹⁰) át kell terelni, integrálni kell a kutatási adatokat a digitális filológiai keretrendszerbe (pl. személynevek – XML-ek automatizált előállítás);
- A kezdetektől össze kell hangolni a párhuzamosan futó projekteket --> *tudás- és adatmegosztás 3. (kutatási adatok terén is).*

Ez utóbbit segíti elő az Entitástár Trello-felülete, amely így a tartalommenedzsment mellett gyakorlatilag egy kutatói kommunikációs platform szerepét is betölti.



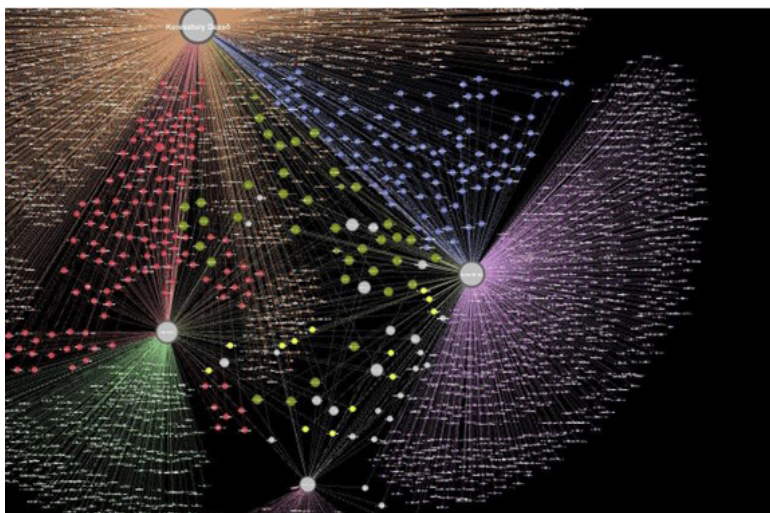
A digitális bölcsész újabb kiemelt (és elhallgatott) feladata tehát a *kooperáció lehetőségeinek megteremtése, illetve az elszigetelt kutatómunkák összekötése.*

5. Szemantikus kapcsolatok, vizualizációk

A digitális kritikai szövegkiadási projektek egyik fontos hozadéka, hogy a strukturált adatok különféle vizuális formákban is megjeleníthetők, ezáltal nemcsak a kutatók számára teszik átláthatóvá az adatok közötti bonyolult relációkat, hanem a szélesebb közönség számára is hozzáférhetőbbé teszik a tudományos eredményeket. A vizualizáció ugyanakkor nem pusztán illusztrációs funkciót tölt be, hanem módszertani eszközként is szolgálhat.

¹⁰ Ez nagyban függ az adott kutató számítástechnikai kompetenciájától, a kutatás mélységétől. Történhet akár a kezdetektől is, de általánosságban inkább a kutatási adatok letisztázása, szakértői validálása után jellemző.

A dHUplán többféle vizualizációs típust alkalmazunk. Ilyen például az írói levelezések térképes megjelenítése¹¹. Az egyes levelezéskiadások keletkező adatai folyamatosan betölthetők erre a felületre, így egy adott helyszínhez különböző időben történő, különböző személyek közötti interakciók (és szövegek) sora kapcsolódik. Más típus a kapcsolati háló, amely a személyek közötti viszonyokat és azok jellemzőit jeleníti meg. Az úgynevezett egohálók¹² (egy személyre koncentráló kapcsolatrendszer feltérképezése) különösen alkalmasak arra, hogy bemutassák az irodalmi élet egy szegmensének, egy adott szerzői kör vagy intézmény szereplőinek viszonyrendszerét, elsősorban a levelezéseik adatai alapján. Az azonos vagy átfedésben lévő korszakba, illetve közegbe tartozó egohálókat összekötve pedig létrejön egy valódi kapcsolati háló, amely már gráf formájában reprezentálja a több csomóponton összekapcsolódó szociális hálókat (*tudás- és adatmegosztás 4.*).

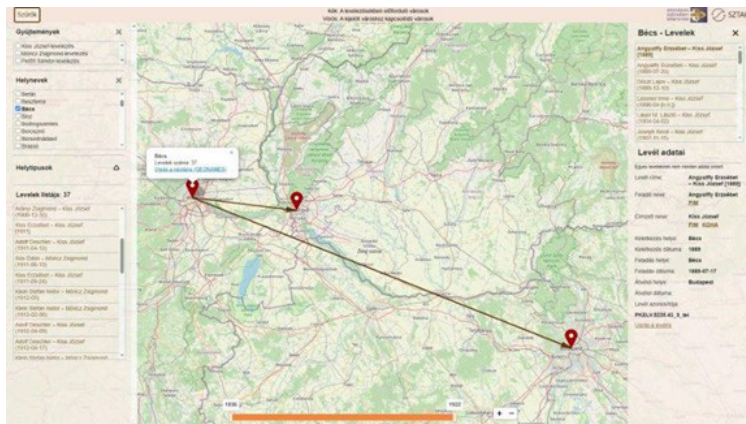


Tervbe van véve a géppel generált kronológia fejlesztése, amely a szövegekhez kapcsolódó és azokban szereplő eseményeket automatikusan kronológiai sorrendben rendezi el, illetve az azonos művekről, alkotásokról szóló szövegek összekapcsolása is. Az RDA (Resource Description and Access) elvén alapuló műentitásháló nemcsak a bibliográfiai kapcsolatok ábrázolására alkalmas, hanem támogatja a szövegtörténeti és recepciótörténeti kutatásokat is.

Kiemelendő, hogy a dHUplán elérhető vizualizációk nagyrészt nem statikus elemek, hanem interaktív módon működnek. Az így megvalósuló vizuális böngészés még inkább lehetőséget teremt új mintázatok felismerésére, valamint a kutatási kérdések újrafogalmazására is.

¹¹ <https://dhupla.hu/page/kreativ/dhupla-map> Hozzáférés: 2025. 08. 07.

¹² <https://dhupla.hu/page/kreativ/vasistvan-kapcsolatihalo> Hozzáférés: 2025. 08. 07.



Konklúzió

A digitális kritikai szövegkiadás nem csupán technológiai innováció, hanem tudomány-szervezési és szemléletbeli fordulat. A strukturált adatok, a szemantikus modellek, az MI-alapú szövegfeldolgozás és egyéb gépi eszközök olyan műhelymunkát tesznek lehetővé, amelyben a digitális filológus központi szerepet játszik: feladata a kritikai szemléletű és tudományos igényességű, továbbhasznosítható, összekapcsolt adathalmazok, szövegek – és emberek kialakítása.

Felhasznált irodalom

- Digital Scholarly Editing: Theories and Practices. Matthew James Driscoll and Elena Pierazzo (eds.), Open Book Publishers, 2016. <https://doi.org/10.11647/OBP.0095>
- Számítógépes irodalomtudomány, Helikon, 2020/1. <https://helikonfolyoirat.hu/helikon-2020-1-szam/>
- Cséve Anna: Levelezés – Móricz-filológia – digitális technológia: Móricz Zsigmond levelezésének digitális kritikai kiadása (2.0 verzió), Szabolcs-Szatmár-Beregi Szemle 57/ 3 pp. 9–15., 7 p. (2022)
- Digital editing and publishing in the twenty-first century. O'Sullivan, J., Pidd, M., Whittle, S., Wessels, B., Kurzmeier, M. and Murphy, Ó. (eds.), Edinburgh, Scottish Universities Press, 2025. <https://doi.org/10.62637/sup.GHST9020>