DOI: 10.31915/NWS.2023.27

Szerkesztői környezet TEI-alapú szövegkiadásokhoz

An Editor Framework for Digital Scholarly Editions in TEI

Mihály Eszter
Országos Széchényi Könyvtár, Digitális Bölcsészeti Központ (OSZK DBK)
mihaly.eszter@oszk.hu

Micsik András
Számítástechnikai és Automatizálási Kutatóintézet, Elosztott Rendszerek Osztály (SZTAKI DSD)
micsik@sztaki.hu

Absztrakt

Az OSZK Digitális Bölcsészeti Központ által fejlesztett új platform, a dHUpla (Digital Humanities Platform – dhupla.hu) elsősorban digitális szövegkiadások publikálására jött létre. Ennek hátterében a szövegkorpuszok előállításához egy teljes szerkesztőségi rendszer kialakítására volt szükség, amely egyrészt felhasználóbarát módon teszi lehetővé a szerkesztést, másrészt funkcióival támogatja a szövegek jelölőnyelvi kódolását. Ezt a felületet a SZTAKI és az OSZK közösen fejleszti egy XML szerkesztő bővítményeként. A szövegkódolás alapja a TEI (Text Encoding Initiative) szabványa, a keretrendszer e nemzetközi ajánlás bonyolult konstrukcióinak bevitelét, többek között kontextus-menüvel, beszúrható mintákkal, Schematron-validációval segíti. A fejlesztésben központi szerepet játszanak a szöveg feldolgozását támogató további eszközök is: névterekkel, adatbázisokkal való összekapcsolás, MI-alapú névelem-felismerés, valamint különböző automatizált műveletek, úgymint PDF- és képkonverzió, vagy adatvizualizáció támogatása. A manuálisan és a gép által végezhető részfolyamatok minden esetben kiegészítik egymást, megteremtve ezzel egy minőségi, a digitális közeg adta lehetőségeket kiaknázó szövegkiadási módszert.

Kulcsszavak: digitális szövegkiadás, TEI, XML szerkesztőfelület

Abstract

The Digital Humanities Centre (DBK) of the Hungarian National Library has developed a platform called dHUpla (Digital Humanities Platform - dhupla.hu) for publishing digitized text editions. For this, the creation of a complete editing environment was necessary to support editors and digital humanists and to help them in the input and validation of correct TEI XML encoding. This environment was implemented as an XML editor extension jointly by DBK and SZTAKI (Insititute for Computer Science and Control). The extension contains custom toolbars, templates, Schematron validation and a set of scripts to automate steps of the conversion to TEI XML. Scripts support named entity recognition and linking, PDF generation, extraction of data for visualizations and other task automations. These operations combine automated execution with manual supervision to reach high quality TEI production.

Keywords: digital scholarly editions, TEI, XML editor

Bevezetés

Az Országos Széchényi Könyvtár Digitális Bölcsészeti Központ (DBK) egyik elsődleges feladata egy korszerű online platform fejlesztése a közgyűjteményekben őrzött szöveges források kezelésére, amely egységes kutatói környezetet jelent az irodalomtudomány, a nyelvtudomány, és más humán tudományok számára. A dHUpla[1] (Digital Humanities Platform) 2021 óta üzemel (2021 decemberéig a Petőfi Irodalmi Múzeum, majd az Országos Széchényi Könyvtár szolgáltatásaként), és számos szövegkiadást tesz nyilvánosan elérhetővé, amelyhez entitástár, valamint kreatív tartalmak, vizualizációk is társulnak. E szolgáltatások hátterében a szövegkorpuszok előállításához egy teljes szerkesztőségi rendszer kialakítására volt szükség, amely egyrészt felhasználóbarát módon teszi lehetővé a szerkesztést, másrészt funkcióival támogatja a szövegek jelölőnyelvi kódolását. Ezt a felületet a SZTAKI és az OSZK közösen fejleszti egy XML-szerkesztő bővítményeként. Az alábbiakban röviden bemutatjuk a dHUpla funkcióit, majd rátérünk a szerkesztői felület részletes ismertetésére.

A dHUpla infrastruktúra

A dHUpla elsődleges célja, hogy a kutatók, olvasók számára egységes felületen, filológiai igényességgel hozzáférhetővé tegye a magyar kulturális örökség különböző intézményekben őrzött, eddig ismeretlen vagy méltatlanul elfeledett szöveges tartalmú, elsősorban kéziratos forrásait. Emellett olyan szerkesztőségi keretrendszert is kínál a tartalmak digitalizálásához, amely biztosítja a források egységes és színvonalas feldolgozását.

A platform alapjait és felépítését korábban már részletesen bemutattuk [1]. A rendszer rugalmas és moduláris, a tartalom előállítása többféle úton történhet. Az átírt szövegek a nemzetközileg támogatott TEI (Text Encoding Initiative) szabvány szerint annotált XML formátumban készülnek, amely a nemzetközi integráció mellett lehetővé teszi többek között a dokumentumok gépi feldolgozását és értelmezését (linked open data), szemantikus hálók kiépítését (semantic web), adatgazdagítást (data enrichment), illetve a távoli olvasás (distant reading) különböző aspektusait. A szolgáltatás kapcsolatot létesít különböző névterekkel, bibliográfiai forrásadatbázisokkal, illetve nyelvi elemző szoftverekkel. Mindezek segítségével a legkülönbözőbb elemzések, korpuszlekérdezések, adatvizualizációk válnak megvalósíthatóvá.

Az infrastruktúra középpontjában a git verziókövető szoftver áll, amely végül teljes mértékben kiváltotta egy XML-adatbázis használatának szükségességét, nagyban leegyszerűsítve a rendszer használatát, karbantartását és fejlesztését. A dHUpla git-ben lévő források (szöveg, programkód) alapján publikál, így a git repository-k birtokában bárhol újraépíthető a teljes dHUpla honlap. A projektek mind önálló git repository-ban vannak, a publikáláshoz szükséges transzformációt docker containerek végzik, minden egyes projekthez meg lehet adni saját ún. buildert, amelyekben tetszőleges programnyelvet lehet használni. A HTML tartalmon túl Apache Solr indexfájlt is előállítunk, amelynek segítségével a legkülönbözőbb facettás keresések is lehetővé válnak.

A publikációs felületen több módon lehetőség nyílik az átírt szöveg és az eredeti facsimile együttes vizsgálatára, a digitális objektumok különböző szempontú rendezésére, szűrések elvégzésére. Az egyes gyűjtemények megjelenésének, funkcióinak konfigurálása egyszerű szöveges fájlokban (yml) történik.

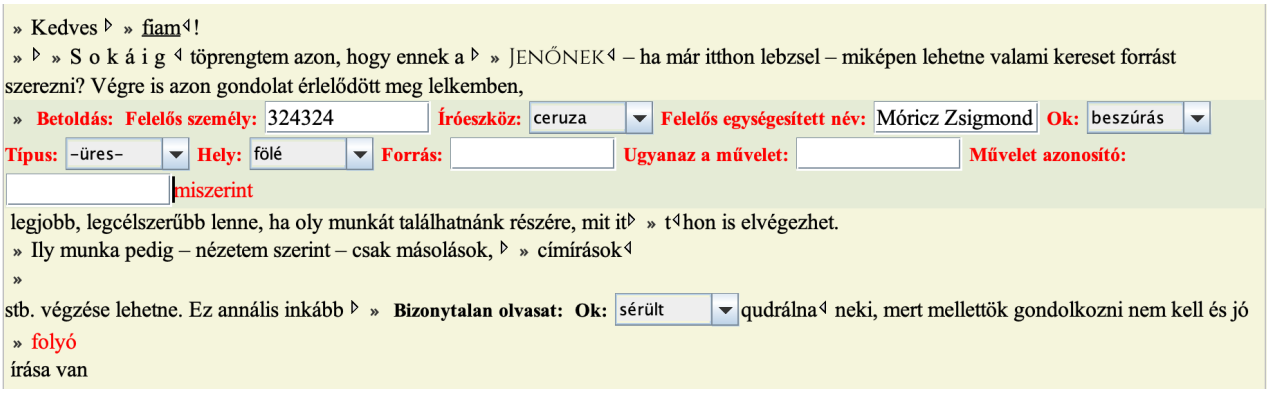
1 dhupla.hu

A kéziratok átírása során ezenkívül olyan kézírásfelismerő-modell épül, amely folyamatosan bővülve alkalmassá válik a magyar nyelvű kézírások automatikus felismertetésére, azaz mesterséges intelligencián alapuló gépi feldolgozására (Handwritten Text Recognition).

A TEI szerkesztői felület

A dHUpla oldalára szánt szövegeket a DBK által készített TEI XML specifikációk szabályai szerint kell létrehozni, mivel a TEI igencsak rugalmas XML formátum, és több különböző ábrázolási mód is választható ugyanarra a szövegelemre. A TEI XML-ek részletes szerkesztésére az Oxygen XML Editor[2] programot használjuk. Az Oxygen XML Editor (röviden továbbiakban Oxygen) beépített TEI támogatással rendelkezik, de ez a már említett teljesen általános, korlátozások nélküli TEI-n alapul. Az Oxygenen belül lehetőség van ún. frameworkök kifejlesztésére, amely egy személyre szabott szerkesztői felületet tud nyújtani, eszköztárakkal, menükkel, specializált megjelenítéssel és ellenőrzési lehetőségekkel.

A frameworkben alkalmazott megoldásokból elsőként magát a szövegszerkesztő ablakot vesszük sorra. Itt háromféle nézetből lehet választani, ebből számunkra kettő a lényeges: az első magát az XML kódolást mutatja, ahol a rideg valóság tárul elénk, és ellenőrizhetjük, illetve változtathatjuk az XML-t. A szerzői nézet ezzel szemben egy CSS-sel barátságossá alakított, olvasható nézetet kínál, ahol az elemek szerkesztése vezetett módon lehetséges (1. ábra).

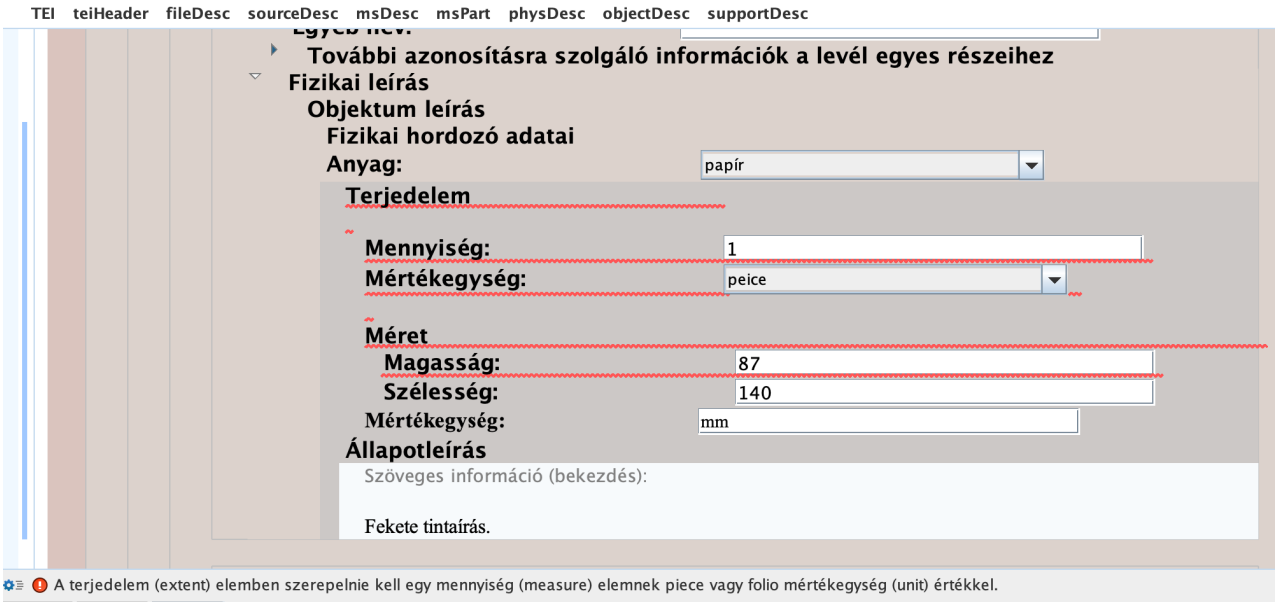


1. ábra: Az ún. szerzői nézetben segítséget kapunk a speciális elemek kitöltéséhez

Az 1. ábrán megfigyelhetjük, hogy a speciális szövegelemek előtt álló » jelre kattintással kinyitható és becsukható az elem tartalma. Az ábrán egy <add> és egy <unclear> elem attribútumait láthatjuk.

A metaadatok beírását részletesen kidolgozott űrlapok segítik (2. ábra). A piros aláhúzás hibás kitöltést jelez, a magyarázat a legalsó sorban olvasható. Az ellenőrzést Schematron[3] szabályokkal végezzük.

2 https://www.oxygenxml.com/xml_editor.html
3 https://www.schematron.com/



2. ábra: A TEI fejléc kitöltése és ellenőrzése

A 3. ábrán láthatóak a szerkesztők munkáját megkönnyítő eszköztárak. A felső sorban található menüvel a gyakori TEI elemeket lehet beilleszteni a szövegbe. A középső sor főleg a fejléc, vagyis a metaadatok kitöltéséhez nyújt segítséget. A legalsó sorból feldolgozási parancsokat lehet indítani. Ebben a sorban nem látszik az összes lehetséges művelet, van még keresés több más névtérben is, illetve PDF készítés mint menüpont.

3. ábra: Az általunk készített eszköztárak

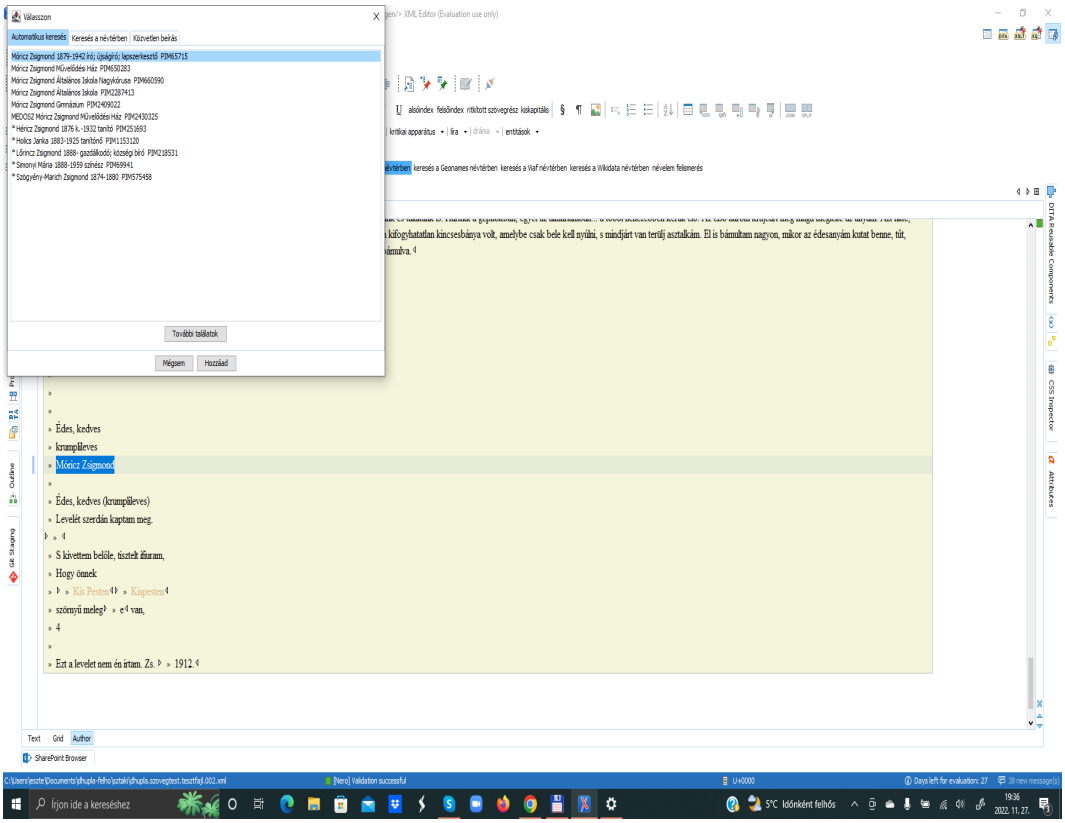
Példa munkafolyamat

A kéziratok kézi és automatikus gépi átírásához, valamint az ún. text-image linking (kép és szöveg összekötése) elvégzéséhez a Transkribus programot használjuk, amely TEI XML formátumban is tud exportálni, azonban ez a formátum nem szabványos, és a DBK előírásainak sem felel meg. Az Oxygen eszköztárból indítható Transkribus export konverzió segítségével azonban megfelelő formára tudjuk hozni a TEI XML-t.

Ezután a szerkesztőben ellenőrizhetjük az XML-t, javíthatjuk a kapott Schematron hibajelzéseket, és egyéb finomításokat tudunk tenni a szövegjelölésekben. Ehhez a fázishoz kapcsolódik a TEI fejléc kitöltése, amelyhez egy teljes sablont tudunk egyetlen gombnyomással beilleszteni, majd a kapott űrlapokat kitölteni.

A szöveg feldolgozásában segít a megemlített névelemek (személyek, települések, szervezetek stb.) bejelölése. Erre saját fejlesztésű névelem-jelölőnket lehet használni, amely a HuSpaCy [2] nyelvi feldolgozó eszköz segítségével beilleszti a megfelelő TEI XML elemeket (pl. <persName>) a talált nevekhez. Ezek a nevek még azonosítatlanok, vagyis nincsenek egy gondozott névtérben a megfelelő névelemhez kötve. Erre a célra az eszköztár keresőgombokat tartalmaz, amely a kijelölt szöveget megkeresi az adott névtérben (pl. PIM[4], GeoNames, Wikidata), és a szerkesztő választásának megfelelően beírja az XML-be a névtér azonosítót (4. ábra).

4 https://opac-nevter.pim.hu/search



4. ábra: Keresés a szövegben kiválasztott névre az Oxygen szerkesztőfelületén

Ha elkészült a TEI XML feldolgozás, az elemzési és terjesztési feladatokat is támogatja a framework. Az XML-ből például PDF-et tudunk előállítani, amely nem tartalmazza ugyan az XML-ben kódolt összes részletet, viszont jól olvasható és megosztható.

Dokumentumok csomagjaiból (pl. egy író levelezése) a Metaadatok kinyerése funkció segítségével táblázatokat kapunk, amelyek aztán különböző összesítések és vizualizációk alapjául szolgálhatnak. Az egyik ilyen vizualizációs eszköz hamarosan elkészül: a levelek feladását és kézhezvételét mutatja térképen.

Összefoglalás

A digitális szövegkiadások TEI alapú publikálása rengeteg új lehetőséget ad a digitális bölcsészeknek, amelyből párat jelen cikkben említettünk. Nehéz viszont odáig eljutni, hogy a szövegek jó minőségű TEI XML formátumba kerüljenek, mivel a folyamat sok szakértelmet és szakértői munkát igényel. Ennek a feldolgozási munkának a megkönnyítésére készítettük el a bemutatott Oxygen XML Editor framework megoldást, amely számos hasznos funkcióval segíti a szerkesztést, de ezen felül lehetőséget teremt a TEI-n belüli saját kódrendszerek megalkotására és az elkészített TEI dokumentumok e szerinti ellenőrzésére, és egy alap-minőség biztosítására a konverziók végén. A jelenlegi framework a levelezések XML kódolására fókuszál, de látjuk már a lehetőséget többféle „szakosodott” leírási mód (pl. drámák, versek, naplók) egyidejű támogatására is. A keretrendszer fejlesztése jelenleg a házon belüli alkalmazás és tesztelés fázisában van.

Irodalomjegyzék

[1] Mihály, Eszter (2022) *Mi az a dHUpla?: A Digitális Bölcsészeti Platform bemutatása.* In: Valós térben - Az online térért, Networkshop 31: országos konferencia. 2022. április 20–22. Debreceni Egyetem. Kiadja a HUNGARNET Egyesület az MTA Könyvtár és Információs Központ közreműködésével, Budapest, pp. 345–358. ISBN 978-615-82243-0-7 DOI: 10.31915/NWS.2022.44

[2] Orosz György, Szántó Zsolt, Berkecz Péter, Szabó, Gergő, Farkas Richárd (2022). HuSpaCy: an industrial-strength Hungarian natural language processing toolkit. In XVIII. Magyar Számítógépes Nyelvészeti Konferencia.

ÚJ TECHNOLÓGIÁKKAL, ÚJ TARTALMAKKAL A JÖVŐ DIGITÁLIS TRANSZFORMÁCIÓJA FELÉ

32. Networkshop: országos konferencia

2023. április 12–14.
Pannon Egyetem, Veszprém

ÚJ TECHNOLÓGIÁKKAL, ÚJ TARTALMAKKAL A JÖVŐ DIGITÁLIS TRANSZFORMÁCIÓJA FELÉ

32. Networkshop: országos konferencia

2023. április 12–14.
Pannon Egyetem, Veszprém

Szerkesztette: Tick József, Kokas Károly, Holl András

HUNGARNET Egyesület
Budapest, 2023

Szerkesztette: Tick József, Kokas Károly, Holl András

Tipográfia és tördelés: Vas Viktória

Networkshop
2023. április 12–14. Pannon Egyetem, Veszprém konferencia előadásainak közleményei

ISBN 978-615-82243-1-4
DOI: 10.31915/NWS.2023

Kiadja a HUNGARNET Egyesület
az MTA Könyvtár és Információs Központ közreműködésével
Budapest
2023

Borítókép: freepik.com

TARTALOMJEGYZÉK