

EGYETEMI KÖNYVTÁR ÉS LEVÉLTÁR 2025. ÉVI LEVÉLTÁRI MŰKÖDÉSI TÁMOGATÁSRÓL

Pályázati azonosító: L-KOZLEV-25-0009

SZAKMAI BESZÁMOLÓ



A Széchenyi István Egyetem (székhely: 9026 Győr, Egyetem tér 1. Képviselő: Dr. Filep Bálint elnök, Nyilvántartási szám: FI38696, Adószám: 19253079-4-08, Csoportos adóazonosító: 17782294-5-08) a Kulturális és Innovációs Minisztérium által hozott támogatási döntés alapján a LEVELTAR-K-25-0009 számú támogató okiratban rögzített 2.100.000 Ft támogatást a Kedvezményezett a támogatási kérelmében, a támogatói okiratban és a költségtervben részletezettek szerint az Albert Kázmér Mosonmagyaróvári Kar elődintézményeihez tartozó jelentős történelmi értékű vezetőtestületi jegyzőkönyvek feldolgozásában, átírásában és nyílt hozzáférésű online közzétételében résztvevő szakemberek személyi kifizetéseire (megbízási díj és járulékai) fordította. A pályázat segítségével egy innovatív, mesterséges intelligenciára (MI) épülő digitalizációs és iratfeldolgozási projektet sikerült megvalósítani.

A levéltár a pályázat megvalósítása során három fő célkategóriát és ahhoz kapcsolódó módszertant határozott meg:

1. **Mesterséges intelligencia alapú szövegátírás:** hét főből álló csapat az 1851–1919 közötti kéziratos, valamint az 1920–1945 közötti gépiratos jegyzőkönyvek feldolgozását és átírását végezte el.

1.1 Kézírást íratok (1851–1909): A Transkribus platformon a csapat saját, magyar nyelvű kézírásfelismerő modelleket tanított be („HunEdu - Council Records” és „Hungarian Higher Education Council Minutes”), amelyekkel kiváló, 4,46–4,94%-os karakterhiba-arányt (CER) ért el. A munkafolyamat során az szövegek átírásához a Google Gemini MI-asszisztent is használták. Külön kihívást jelentett az 500 oldalas, német nyelvű Kurrentschrift szöveg javítása, amely a speciális földrajzi és személynevek, valamint a szakkifejezések miatt a vártnál több nehézséget okozott.

1.2 Gépiratos iratok (1909–1949): Mivel a hagyományos eszközök (mint az ABBYY FineReader) a halvány és kopott, egyedi struktúrájú gépiratok miatt nem elég hatékonyak, a gépiratok átírása Transkribusban túlzottan drága lett volna, a kipróbált Kraken rendszer pedig rugalmatlannak bizonyult, a levéltár saját, nyílt forráskódú megoldásokat használó Python-alapú munkafolyamatot épített.

A képeket Scantailorral előfeldolgozták, a szövegszegmentálást YOLO v11n modellel, míg magát a karakterfelismerést a Microsoft TrOCR segítségével végezték. A nagy számításigényű modellek futtatásához a Kaggle.com felhős platformját és GPU gyorsítóit használták, amivel a feldolgozási időt a töredékére csökkentették.

A projekt időtartama alatt, a technológia rendkívül gyors fejlődését látva a levéltár úgy döntött, hogy a teljes gépiratos állomány karakterpontos, manuális javítása helyett a meglévő tanítóanyagok (Ground Truth) és az XML-kép párosok hosszú távú megőrzésére fókuszál, így a jövőben az újabb, fejlettebb modellekkel újra felismertethető lesz a szöveg.

Segédletek, jegyzékek készítése és adatbázis-építés: A jegyzőkönyvek feldolgozásával párhuzamosan elkészítésre került az 1851 és 1907 közötti időszak 320 ülésének jegyzéke, valamint az 1920 és 1949 közötti időszak 432 ülésének részletes napirendi jegyzéke, amely mintegy 10 200 napirendi pontot foglal magába. A technológiai innováció itt is megjelent: a kinyert szövegeket RAG technológiával (Gemini Notebook) elemezték, hogy az üléslistákból

automatikusan Python szkript segítségével PDF-könyvjelzőket generáljanak a 13,5 GB méretű, 41 fájlból álló végleges állományhoz.

- Tömeges digitalizálás és karakterfelismerés:** Az 1970–1999 közötti jegyzőkönyveket saját eszközökkel (Canon dokumentumszkennerrel) 300 dpi felbontású PDF formátumban digitalizálták. Összesen 15 600 oldal digitális másolata készült el. A technológia rendkívül gyors fejlődését látva a levéltár úgy döntött, hogy ehhez az anyaghoz új modelleket tanít be: a bevált YOLO+TROCR pipeline mellett a Microsoft Florence-2 VLM modelljét is kipróbálja. A 400 oldalas tanítóanyag előállítása zajlik.

Összegzés és intézményi hatások: A projekt szorosan illeszkedik a Levéltár 2025–2030 közötti stratégiájának állományvédelmi és online szolgáltatási céljaihoz. A feldolgozott, digitálisan kereshető és könyvjelzőzött vezetőtestületi jegyzőkönyvek (eddig 6291 oldalnyi anyag) nyílt hozzáféréssel elérhetővé válnak a levéltár online repozitóriumban (<https://rep.szerep.sze.hu/home>), és tervezik más platformon való közzétételüket is. Emellett az XML formátumban megőrzött adatok a jövőben lehetővé teszik a fejlett NLP (természetes nyelvfeldolgozás) és NER (név- és kinyerés) alapú adatgazdagítást, illetve saját RAG technológia kidolgozását is. A tanácsulási jegyzőkönyvek a legfontosabb intézménytörténeti források. A projekt ezeket a dokumentumokat teszi modern formában, bárki számára online kutathatóvá.

Új magyar kézírásfelismerő (HTR) modell: A „Hungarian Higher Education Council Minutes” (4,46% hibaarány) néven a publikált modell az Országos Széchényi Könyvtár (OSZK) modellje után a második magyar nyelvű, folyó szövegen tanított modell a Transkribus platformon.

A HTR modell országos szintű alkalmazhatósága: A 323 ezer szavas magyar tanítókorpuszon alapuló, a későbbiekben előállítandó kézírásmodell kiváló alapot biztosít majd más hazai egyetemek tanácsulási jegyzőkönyveinek átírásához is, mivel kevés saját adattal finomhangolva náluk is hatékonyan működhet. Az OSZK hasonló nagyságrendet képviselő átíratával összevonva pedig még szélesebb körű felhasználást tesz lehetővé a jövőben.

A digitalizációs projekt közvetlenül támogatja az egyetem hírnév-kommunikációját, az intézményi – kiemelten a több mint 200 éves mosonmagyaróvári agrár szakképzéshez kapcsolódó – örökség megőrzését, továbbá mivel az iskolai közösségi szolgálatos középiskolásoknak is bemutatták a kidolgozott folyamatokat, a hallgatótoborzáshoz is tevőlegesen hozzájárul. A projekthez kapcsolódóan a részt vevő kollégák számára két workshopot, a mesterséges intelligencia alapú feldolgozórendszer első verziójáról szakmai közönség számára egy előadást tartottak a Debrecenben 2026 áprilisában rendezett Networkshop konferencián. A levéltár az informatikai és digitalizációs tapasztalatokat szívesen megosztja a Magyar Felsőoktatási Levéltári Szövetség tagjaival, segítve a hasonló méretű intézmények levéltári munkáját.

Győr, 2026. július 30.

Tisztelettel és köszönettel:



Tóth Csilla

EKL igazgató

