

Jelentős erdélyi sajtóanyag digitalizálása valósult meg ELTE és a Digitéka együttműködésében

Sikeresen zárult az Eötvös Loránd Tudományegyetem konzorciumi vezetésével működő Digitális Örökség Nemzeti Laboratórium ([DH-LAB](#)) és az Erdélyi Digitális Tudománytár ([Digitéka](#)) közös projektje, amelynek keretében több százezer oldalnyi erdélyi magyar sajtóanyag vált kereshetővé és hosszú távon megőrizhetővé korszerű digitális technológiák alkalmazásával.

A felek célja az volt, hogy a történeti erdélyi sajtóforrások digitális feldolgozását új szintre emeljék, és egyúttal hozzájáruljanak a magyar nyelvű kulturális örökség korszerű kutathatóságához.

A munka első szakaszában **26 erdélyi történeti újság mintegy 273 000 oldalnyi képfelvételén** hajtották végre az optikai karakterfelismerést (OCR). Ezt követően további több mint 60 000 oldalt küldött a partner. Így összesen **333 492 oldal** magyar nyelvű erdélyi sajtóanyag feldolgozása készült el. Az elkészült állományokat a Digitéka számára kétrétegű, kereshető PDF formátumban, egységes vízjellel ellátva adták át.

A projekt szakmai jelentősége túlmutat a digitalizáláson. Az OCR-folyamatok hatékonyságának növelése érdekében a felek – az ELTE kutatási és fejlesztési kompetenciáira és infrastruktúrájára is támaszkodva – közösen fejlesztették a dokumentumok szerkezeti felismerését végző, úgynevezett layout analysis rendszert. Ennek keretében az Digitéka annotátorai 1007 oldalt dolgoztak fel, amely a DH-LAB annotátorai által feldolgozott anyaggal együtt összesen 4078 annotált oldalból álló tanítóadatbázist eredményezett. Ez az adatállomány egy olyan, kifejezetten az erdélyi és magyar történeti dokumentumokra optimalizált layout-felismerő rendszer kialakítását alapozza meg, amely számottevően javítja az OCR eredmények pontosságát.

Sajtókapcsolat:

- kommunikacio@elte.hu

Eredeti tartalom: Eötvös Loránd Tudományegyetem

Továbbította: Helló Sajtó! Üzleti Sajtószolgálat

Ez a sajtóközlemény a következő linken érhető el:

<https://hellosajto.hu/?p=28908>