DOI: 10.31915/NWS.2021.10

A digitális szövegkiadások nehézségei és lehetőségei a közgyűjteményekben

Mihály Eszter
Petőfi Irodalmi Múzeum Digitális Bölcsészeti Központ
mihaly.eszter@dbk.pim.hu

Közgyűjteményi feladatok kéziratos források esetén

A kéziratos forrásokat őrző közgyűjtemények hagyományos feladatait a digitális technológiák fejlődése és rohamos elterjedése miatt egyre sürgetőbb újragondolni, átformálni, kibővíteni. A hagyományos eljárások, mint a források leírása, metaadatolása, gyűjteményezése már nem látszik elegendőnek, illetve ezek módszere is korszerűsítésre szorul. Az archiválás, hosszútávú megőrzés, illetve a kereshetővé és hozzáférhetővé tétel szintén mást jelent digitális közegben, mint az analóg világban.

A források digitális másolatainak létrehozásának általános gyakorlatával az is kérdésessé vált, hogy a közzététel mely szintjét nevezhetjük már publikációnak. Ha egy kéziratos forrás digitális másolatát közzétesszük, az publikációnak számít-e?[1]

Mindenesetre ma már alapvető feladatnak tekinthetjük, hogy a közintézményekben őrzött kéziratos források valamilyen módon hozzáférhetőek legyenek anélkül, hogy a fizikai példányt a kezünkbe kéne vennünk. Ez viszont azt jelenti, hogy a szövegkiadás legalább is bizonyos szintjei a közgyűjtemények hatáskörébe kerültek. A GLAM szektor kéziratos forrásokat őrző szereplőinek ezen új feladatkör kihívásaival kell szembesülniük, és válaszokat adniuk.

A kéziratos források közzététele több szinten lehetséges:

1. *A kéziratok digitális másolatának közzététele*
   Mivel az archiváláshoz is mindenképp szükséges egy jó minőségű digitális fakszimile készítése, nem jelent sok többletmunkát az ún. master fájlból egy közzétételre alkalmas (kisebb felbontású, de jól használható) verzió létrehozása, amely automatikus konverzióval megoldható. A szkennelés elvégzésén túl csak az intézmény online katalógusfelületére való feltöltés elvégzésére van szükség (és természetesen előzetesen az adott forrás gyűjteményezésére).

   Így a gyűjteményi rekord adatai, tehát a kéziratok metaadatai kereshetővé, szűrhetővé válnak (az adott katalógusfelület lehetőségei szerint), a felhasználó pedig azonnal megkapja olvasásra a kézirat digitális másolatát.

   Ez a szint az analóg világban a hasonmás kiadásnak feleltethető meg, bár az online keresőfelület ehhez számos egyéb funkciót ad hozzá.

2. *A kéziratok digitális másolatának és átiratainak közzététele*
   Az előző szinthez képest ez az eljárás egy lépéssel továbbmegy, nem csak digitális fakszimiléket készít, hanem a szövegek átírását is elvégzi. Ez természetesen idő-

1 Erre a kérdésre itt nincs lehetőség részletesebben kitérni, a téma külön tanulmány tárgyát képezi.

és erőforrásigényes munka, amelyre nem biztos, hogy mindenhol van kapacitás. A jövőben viszont a nyomtatott szövegek automatikus gépi felismertetéséhez (OCR) hasonlóan egyre több lehetőség nyílik a kéziratos dokumentumok gépi szövegfelismertetésére is (ld. később), amely nagyban csökkenteni fogja a digitális szövegek létrehozásához szükséges erőforrásigényeket.

Az eredmény pedig nagy kilátásokat rejt magában, hiszen a kéziratok teljes szövegállománya kereshetővé válik, nem csak a metaadataik, amely a kutatók, felhasználók előtt hihetetlen távlatokat nyit meg, illetve létrejön az alapvető nyersanyag egy digitális forráskiadáshoz. (Ezen a szinten készülhet betűhű, betű szerinti, illetve normalizált átirat is a projekt céljainak megfelelően.)

Az átírt szöveg és a digitális fakszimile együttes megjelenítésére célszerű formátum a kétrétegű PDF, ahol a felhasználó az eredeti kézirat fotóját látva kereshet a kép mögött rejtve tárolt szövegátiratokban. Ehhez szükséges a szkennelt kéziratképek ún. szegmentálása is, amelynek során zónákra osztjuk a képet, s a zónák koordinátáinak használatával lesz összekötve a kép az átírt szöveggel.

A kétrétegű PDF-ek szintén feltölthetőek az adott intézmény meglévő online katalógusfelületére az előző pontban leírtak szerint, s itt kell lehetőséget nyújtani az ún. teljes szövegű, tehát a gyűjteményi rekordokhoz kapcsolt PDF-ek tartalmában való keresésre[2].

3. *Digitális forráskiadás*
   Ezen a szinten hozhatunk létre egy textológiai-filológiai szempontból forráskiadásnak tekinthető publikációt. Az előző pontban született nyersanyagon tovább munkálkodva, vagy azt kihagyva és újonnan létrehozva születhet meg a kézirat annotált, adatgazdagított, részletesen metaadatolt kiadása. Erre a célra alkalmas a TEI XML[3] formátum, amely nemzetközi szinten is a szövegek annotálására leginkább elfogadott és elterjedt jelölőnyelv. Az ún. címkék (tagek) és tulajdonságaik (attribútumok) használatával lehetőség nyílik szintaktikai (pl. bekezdés) és szemantikai (pl. név) szövegegységek beazonosítására, megjelölésére, ennélfogva lekérdezhetővé, kereshetővé és különböző módokon megjeleníthetővé válnak. A TEI fejlécben (TEI Header) részletes metaadatokat rögzíthetünk, illetve szerkesztői megjegyzésekkel, külső adattárakra való hivatkozásokkal gazdagíthatjuk a szövegközlést. A TEI XML publikációjára külön platform kialakítása szükséges[4], amely kiaknázza az annotált, részletesen feldolgozott szövegekben rejlő megjelenítési, keresési és egyéb lehetőségeket.

2 Természetesen lehet átírást készíteni egyéb formátumokban is, pl. sima szövegként (txt), Word-ben, de figyelembe kell vennünk, melyek a kitűzött céljaink, szeretnénk-e a digitális fakszimilét összekapcsolni az átírt szöveggel, illetve kívánunk-e bizonyos alapvető textológiai jelenségeket már ezen a szinten is jelölni, annotálni (pl. törlések).

3 Text Encoding Initiative: https://tei-c.org/

4 Illetve megoldást nyújthat egy közösségi platform használata. Többek között erre a célra tervezzük a Digital Humanities Platform (dHUpla) felületét.

A digitális forráskiadás textológiai-filológiai munkájának mélysége természetesen nagyon különböző mértékű lehet, a projekt céljainak megfelelően kell meghatározni, sőt a feldolgozást további egymásra épülő fázisokra lehet bontani (pl. az alapvető textológiai annotációk elhelyezésétől a tulajdonnevek beazonosításáig). Így már az első fázis elvégzésével a szövegek publikálhatóvá válnak, és a további feldolgozás során csak frissíteni kell őket.

A forráskiadás alapját képezheti egy későbbi kritikai (tudományos[5]) kiadásnak is, amely már nem feltétlenül a forráskiadást készítő intézmény feladatkörébe tartozik, de adott feltételek mellett természetesen ott is megoldható.

Minden GLAM szektorhoz tartozó intézménynek fel kell mérnie a saját lehetőségeit, kapacitását, és ahhoz mérten választani a fenti lehetőségek közül.

Miért jó a digitális szövegkiadás?

A digitális szövegkiadásoknak tulajdonképpen ugyanazok az elsődleges előnyei, amelyek a hátrányai is. Rugalmasak, ennélfogva állandóan változnak, így folyamatosan javíthatóak is. Esetükben nincs terjedelemhatár, bármeddig bővíthetőek, tehát nem szükséges válogatásokat eszközölni - hacsak nem humánerőforrás-hiány miatt.

Abszolút előnyük viszont az analóg kiadásokhoz képest, hogy annotálhatóak, összeköthetőek egyéb tudástárakkal (pl. névterek, bibliográfiák, tezauruszok), több formában és folyamatosan publikálhatóak, verziózhatóak, illetve a legkülönfélébb műveletek végezhetők bennük (pl. keresés, szűrés, adatvizualizáció). Egészen új módszerekkel válnak kutathatóvá a szöveges tartalmak (ld. “distant reading”[6]), illetve az eredeti források archiválására is alkalmasak.

A digitális szövegkiadások ettől függetlenül nem kell, hogy kizárják a nyomtatott szövegkiadások lehetőségét, sőt. A tendenciák inkább azt mutatják, hogy nagyon is tudnak egymás mellett létezni, más-más felhasználói igényeket kielégítve. A jövő afelé halad, hogy a digitális szövegkiadások lehetnek a nyomtatott kiadások alapjai, s nem fordítva.

Új feladatok, problémák

Milyen új feladatokkal, problémákkal kell megküzdenie egy közgyűjteménynek, ha – valamilyen formában – vállalja a kéziratos forrásai közlését?

Először is össze kell hangolnia a digitalizálást végző műhely, a kéziratokat őrző gyűjteményi tár munkáját, illetve a digitális bölcsészeti feladatokat. Minden esetben meg kell találnia a közös nevezőt, amely az összes fél igényeinek megfelel. Ez korántsem

5 *A kritikai kiadás*, illetve tudományos kiadás (digital scholarly edition) terminológiai kérdései szintén egy külön tanulmány tárgyát képezik.

6 https://en.wikipedia.org/wiki/Distant_reading

egyszerű, már csak a meglévő eszközök, hosszú idők során kialakult hagyományok és protokollok fényében sem[7].

Alapvetően a közgyűjteményi és a digitális bölcsészeti szempontok egyeztetéséről van szó, amely elengedhetetlen a közös munka elvégzéséhez. Ha egy közgyűjtemény kéziratos források közzétételére vállalkozik, elkerülhetetlenül beengedi a digitális bölcsészeti feladatokat az intézmény hatáskörébe, és integrálnia kell azokat.

Ugyanígy, bizonyos digitális bölcsészeti eszközöket is be kell építenie a meglévő közgyűjteményi infrastruktúrába, illetve adott esetben visszafelé, egyes digitális bölcsészeti eszközökbe integrálhatja a már meglévő közgyűjteményi infrastruktúra elemeit (ld. később).

Kritikus probléma a felmerülő többletfeladatokhoz szükséges humánerőforrás biztosítása (pl. szövegek átírása). Amennyiben van rá lehetőség, mindenképpen kézenfekvő és célravezető a kéziratokat őrző gyűjteményi tár dolgozóinak betanítása, és bevonása a digitális bölcsészeti feladatokba. Ezenkívül feltétlenül szükséges az intézmény különböző területein dolgozó informatikusokkal való folyamatos együttműködés is.

Pilot projekt

A fentiekben felvázoltak miatt a Petőfi Irodalmi Múzeum Digitális Bölcsészeti Központjában működő Digitális bölcsészet munkacsoporttal elindítottunk egy pilot projektet, ahol igyekeztünk a felmerülő kérdésekre választ adni, módszereket, protokollokat, munkafolyamatokat kialakítani, megfelelő eszközöket találni, a digitális forráskiadás architektúrájához teszt-modulokat fejleszteni. Az alábbiakban ezeket fogom vázlatosan bemutatni.

A pilot projekt Kiss József[8] levelezésének digitális forráskiadása.

7 Jó példa levelezéskiadás esetén a digitális objektum egységének meghatározása. A PIM Kézirattár a gyűjteményezés során általában teljes levelezési irányokat ír le, amely egy személy másik személynek írott összes levelének metaadatait tartalmazza. A szövegfeldolgozás során az alapegységnek viszont nyilvánvalóan a levélnek kell lennie. Össze kellett tehát hangolni a szkenneléstől a kiadásig, mikor mi legyen az alapegység, a létrejött digitális objektumokat hogyan kapcsoljuk össze a gyűjteményi rekordokkal, és hogyan jelenítsük meg azokat.

8 Kiss József (1843-1921), A Hét című hetilap alapítója „olyan szerkesztői tehetség, akit csak az egy Osvát Ernő múlt felül. Lapja rövidesen orgánuma lett mindazoknak, akikben az új polgári szellem irodalmi formát öltött, de hordozta a régebbi nemzedékek kritikusabb elméit is." – írja Szerb Antal [*Szerb Antal: A magyar irodalom története. Budapest, Magvető, 1972 (ötödik kiadás), 420. o*]. A Hét című folyóirat a Nyugat elődjének tekinthető, amelyre az irodalomtörténeti kutatások eddig méltatlanul kis figyelmet fordítottak. Kiss József szerkesztőségi és egyéb levelezésének közzététele eddig feltáratlan közeget tesz hozzáférhetővé mind a szakmai közönség, mind az olvasók számára.

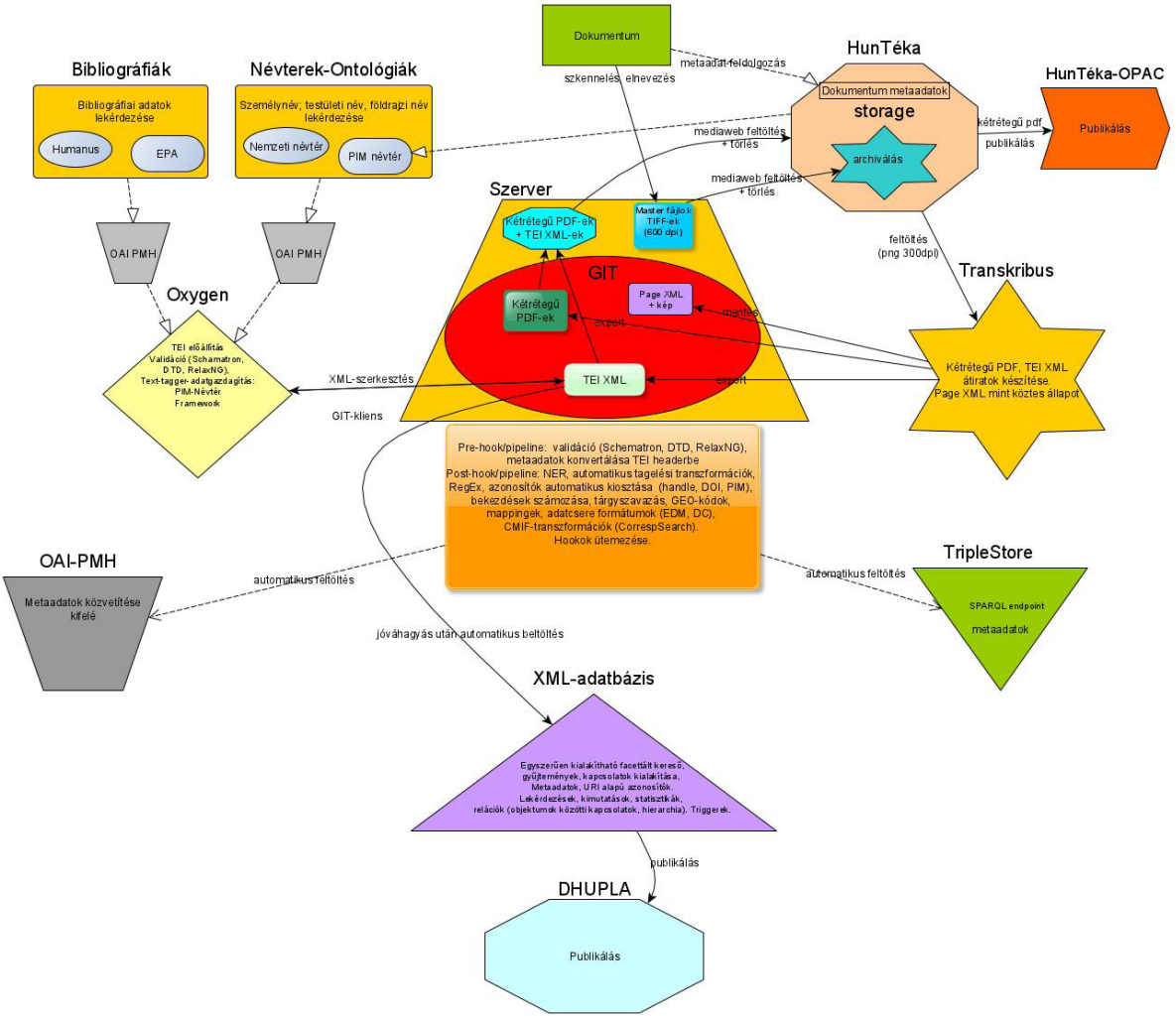
Első lépések

Megtettük a szükséges első lépéseket:

- szkennelés szabályainak kialakítása
- névkonvenció meghatározása (fájlok elnevezése, strukturálása)
- nyilvántartások, kimutatások készítése a projekt forrásairól
- ütemezés
- content management környezet kialakítása
- eszközök kiválasztása
- szerepek kiosztása
- workflow megtervezése

Workflow, infrastruktúra

A kialakított infrastruktúra, s ez alapján a munkafolyamat egyes lépéseinek meghatározása a következő ábrán látható:



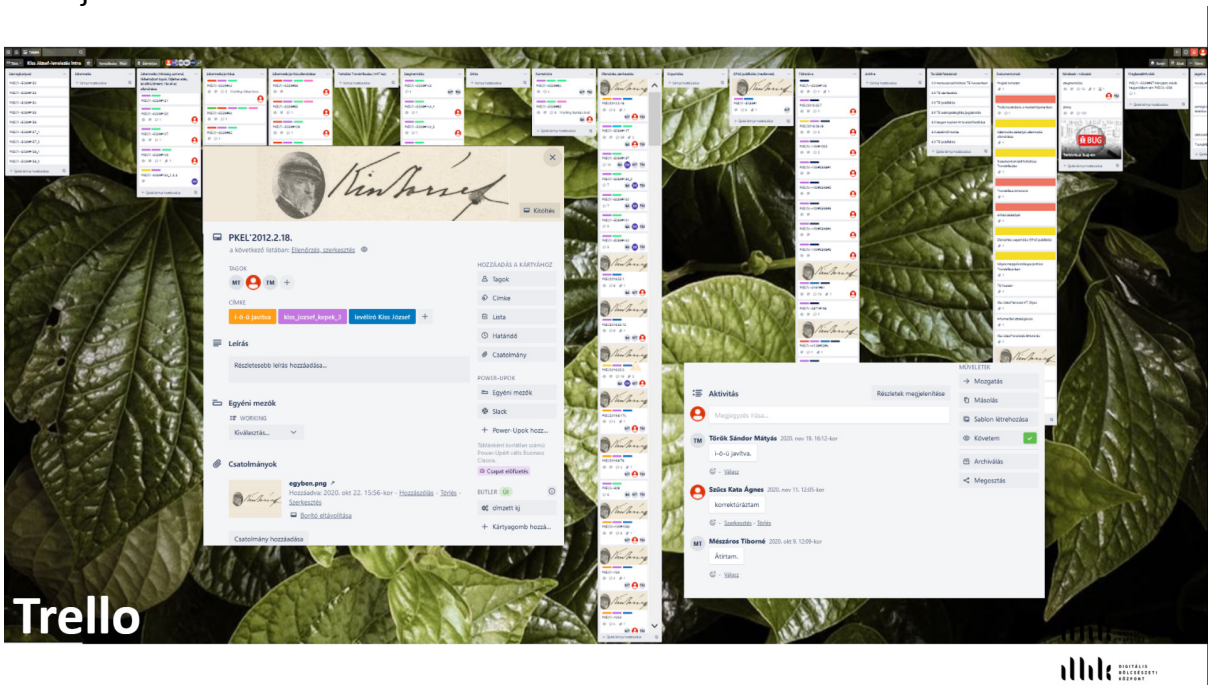
A szkennelés-szövegátírás-publikálás egyszerű hármasa helyett láthatóan sokkal összetettebb rendszerre volt szükség, ugyanakkor mégsem olyan bonyolult, mint amilyennek első ránézésre tűnik. A középpontban egy verziókövető rendszer áll,

amelyben a tárolt anyagok minden változása visszakövethető, s innen indul ki minden automatizált folyamat, ezzel áll kapcsolatban közvetlenül vagy közvetve az összes használt eszköz.

Kétszintű szövegközlést tűztünk ki célul, előbb kétrétegű PDF-ek formájában a PIM Opac felületén[9], majd TEI XML-ek formájában egy új, erre a célra fejlesztett platformon[10], amely az első fejezet 2-es és 3-as pontjának felelnek meg. (A TEI XML-ek publikációja több fázisra van osztva a feldolgozás mélységének megfelelően, hogy minél előbb publikussá válhassanak a szövegek. A névterekkel, bibliográfiákkal való összekapcsolás például csak a harmadik fázisban történik.) Beépítettük a rendszerbe a metaadatok automatikus továbbításának lehetőségeit, illetve a különböző lekérdezési módszerekhez szükséges eszközöket is.

Content management

Az egyes dokumentumok munkafolyamatban történő haladását, a Trello[11] felületén követjük.



Az ún. Kanban-módszer[12] testreszabása eredményeképpen minden oszlop a workflow egy állomásának felel meg, s minden dokumentumnak van egy külön kártyája. A kártya tartalmazza az összes szükséges információt: ki dolgozott már a dokumentummal, jelenleg épp ki dolgozik rajta, illetve minden egyéb a projekt szempontjából szükséges metaadatot, amelyek szűrhetőek is. Ezenkívül hozzászólások formájában lehet megjegyzéseket írni, kérdéseket megvitatni az adott dokumentummal kapcsolatban.

9 https://resolver.pim.hu/gyujtemeny/levelek/media/csatolt/"Kiss József 1843-1921"

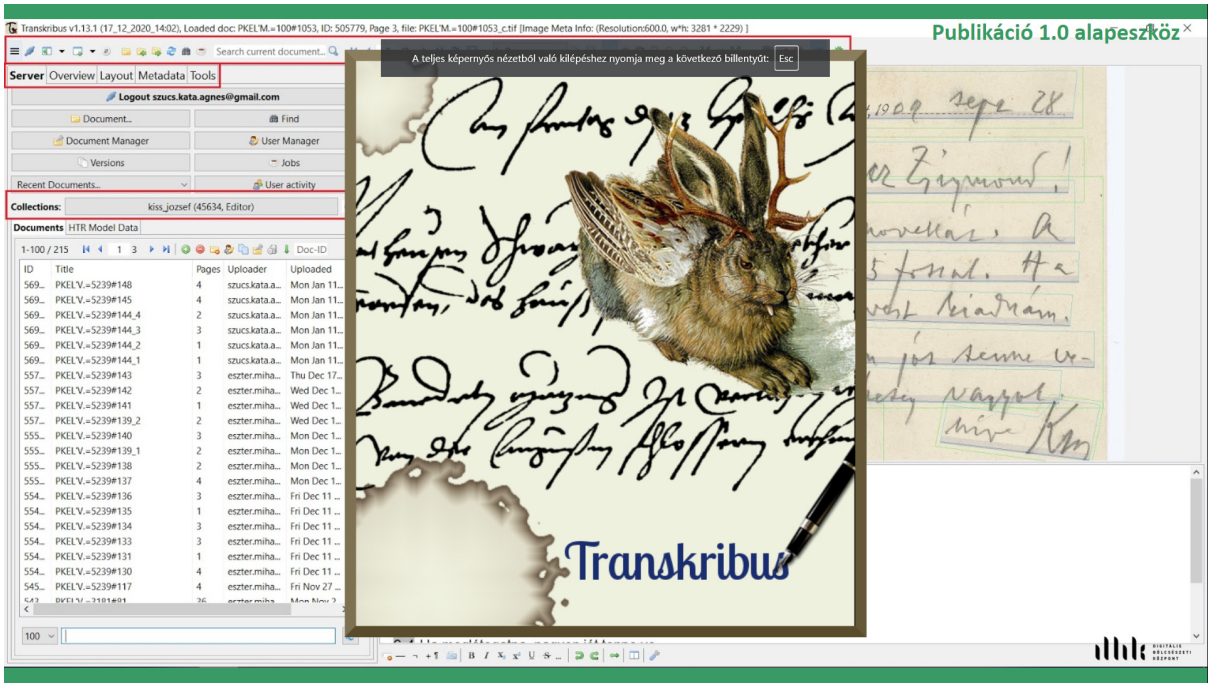
10 dHUpla: Digital Humanities Platform - hamarosan elérhető a demo a www.dhupla.hu oldalon

11 https://trello.com/

12 https://en.wikipedia.org/wiki/Kanban

Publikáció 1.0 - kétrétegű PDF

Szerkesztő program
A kétrétegű PDF fájlok előállításához a Transkribus[13] programot használjuk.



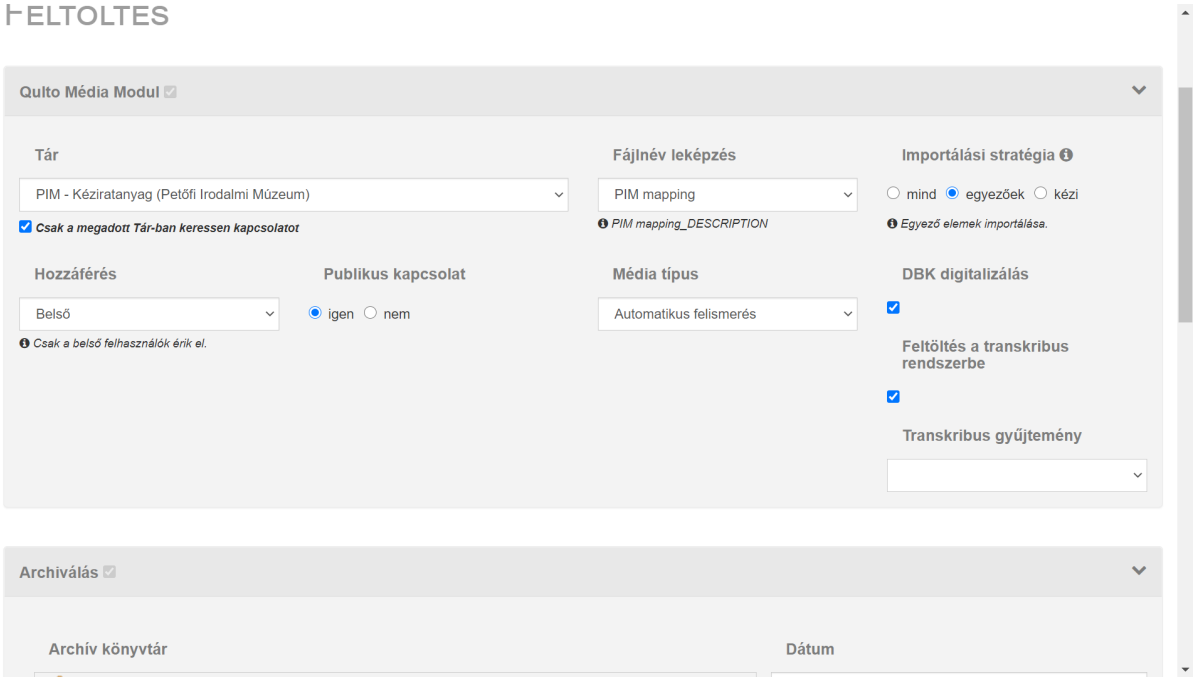
A programot kifejezetten kéziratos dokumentumok átírására fejlesztették, felhasználóbarát felülettel rendelkezik, s nem mellesleg automatikus kézírásfelismertető, ún. HTR (Handwritten Text Recognition) modellek létrehozására is alkalmas. Az átírt dokumentumok többféle formátumban exportálhatóak, mint például a PDF és a TEI XML.

A programmal lehetséges a saját gépen való szerkesztés is, de az összes funkció kihasználásához a Transkribus szerverére is fel lehet tölteni a dokumentumokat, amely szintén verziókezelővel biztosítja a tárolásukat.

Eszközök integrációja
A számos felhasználói funkció teljes igénybevétele miatt a projekt során a szerveren dolgozunk, ezért egyik modulfejlesztésként a Transkribus-gyűjtemények automatikus napi mentésének megoldását tűztük ki célul, amely mára meg is valósult: a PIM saját szerverének verziókezelő rendszerébe naponta mentés készül a Transkribus szerveréről.

A Transkribus programot integráltuk a meglévő közgyűjteményi infrastruktúrába is: az ún. Mediaweb eszközbe, amely eddig a gyűjteményi rekordokhoz tartozó médiatartalmak feltöltésére szolgált, beépítettük a Transkribus-ba való feltöltés funkcióját is.

13 https://readcoop.eu/transkribus/?sc=Transkribus



Ezzel jelentősen leegyszerűsödött a munkafolyamat kezdeti szakasza, tudniillik az archiváláshoz előállított szkennelt kéziratfotók gyűjteménykezelő-rendszerbe (Huntéka) való feltöltésekor a digitális fakszimilék egy gombnyomással feltölthetők a Transkribusba is (a rendszer automatikusan konvertálja azokat a megfelelő felbontásra és formátumba), s azonnal megkezdődhet az átírásuk.

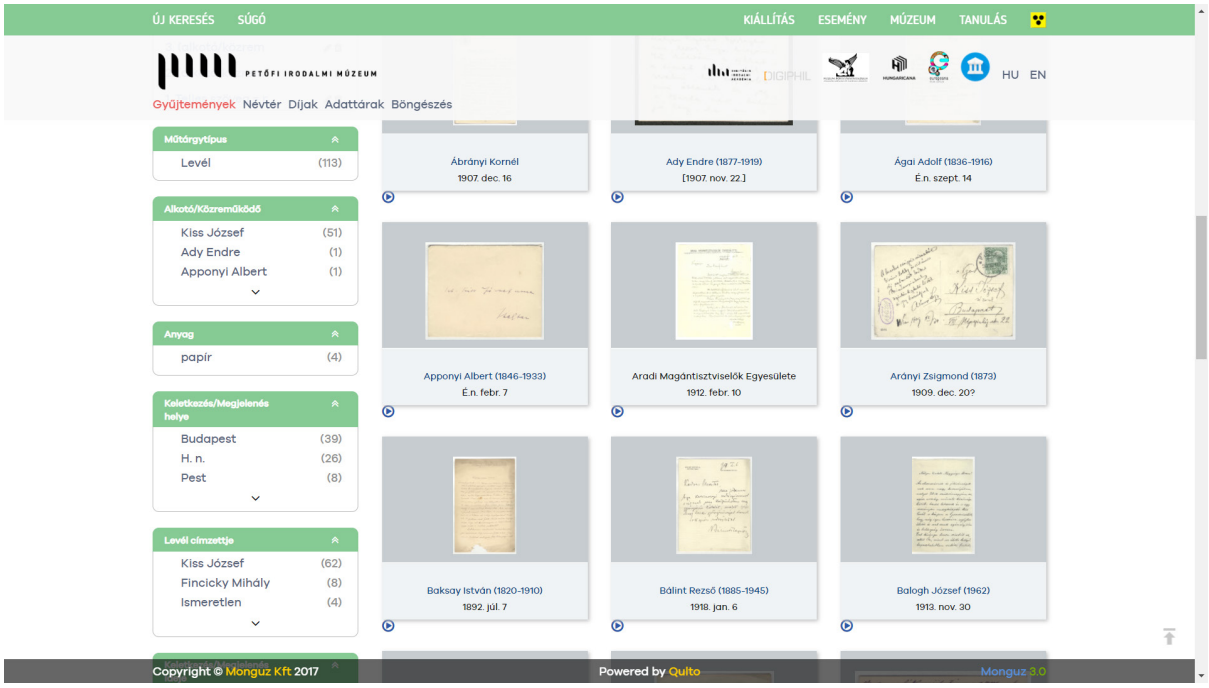
Publikációs, szűrési, keresési, rendezési lehetőségek

A kétrétegű PDF-ek publikálásával viszonylag gyors munkafolyamat valósult meg, az Opac-on kapcsolt médiarekordként való közzététel szinte semmilyen fejlesztést nem igényelt. (A PDF-ek megjelenítésével és a találati listákkal kapcsolatban merültek fel kisebb testreszabás-jellegű feladatok.) Bebizonyosodott tehát, hogy ebben a formátumban jól megoldható a szövegközlés integrációja a közgyűjteményi rendszerbe. A PIM Opac-ján külön URL-en elérhető a teljes Kiss József-levelezés[14], és a levelezés azon része, amely már PDF formátumban is hozzáférhető[15].

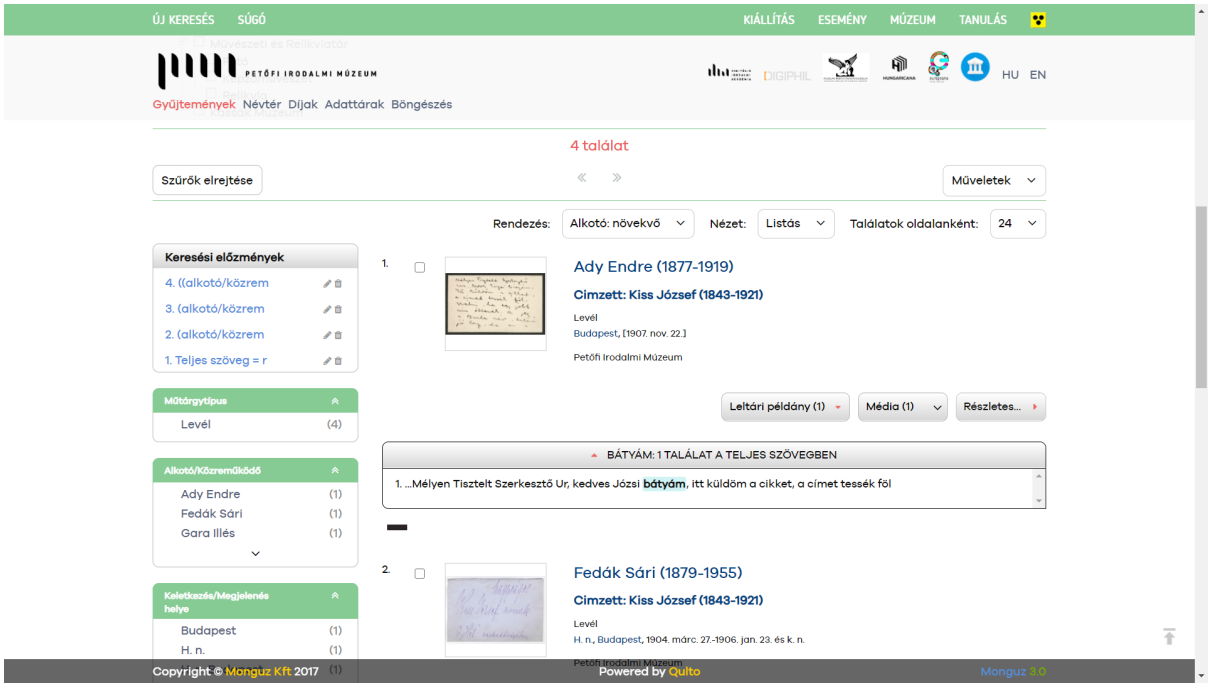
Az online katalógusfelületen eddig is lehetőség nyílt teljes szövegű keresésre, illetve metaadatok szerinti szűrésre is. Az ún. facetek segítségével (ld. alábbi képen bal oldalt) szűkíthetjük a találatokat az egyes adatmezők szerint, és a találati listánkat különböző szempontok szerint rendezhetjük is.

14 https://resolver.pim.hu/gyujtemeny/levelek/"Kiss József 1843-1921"

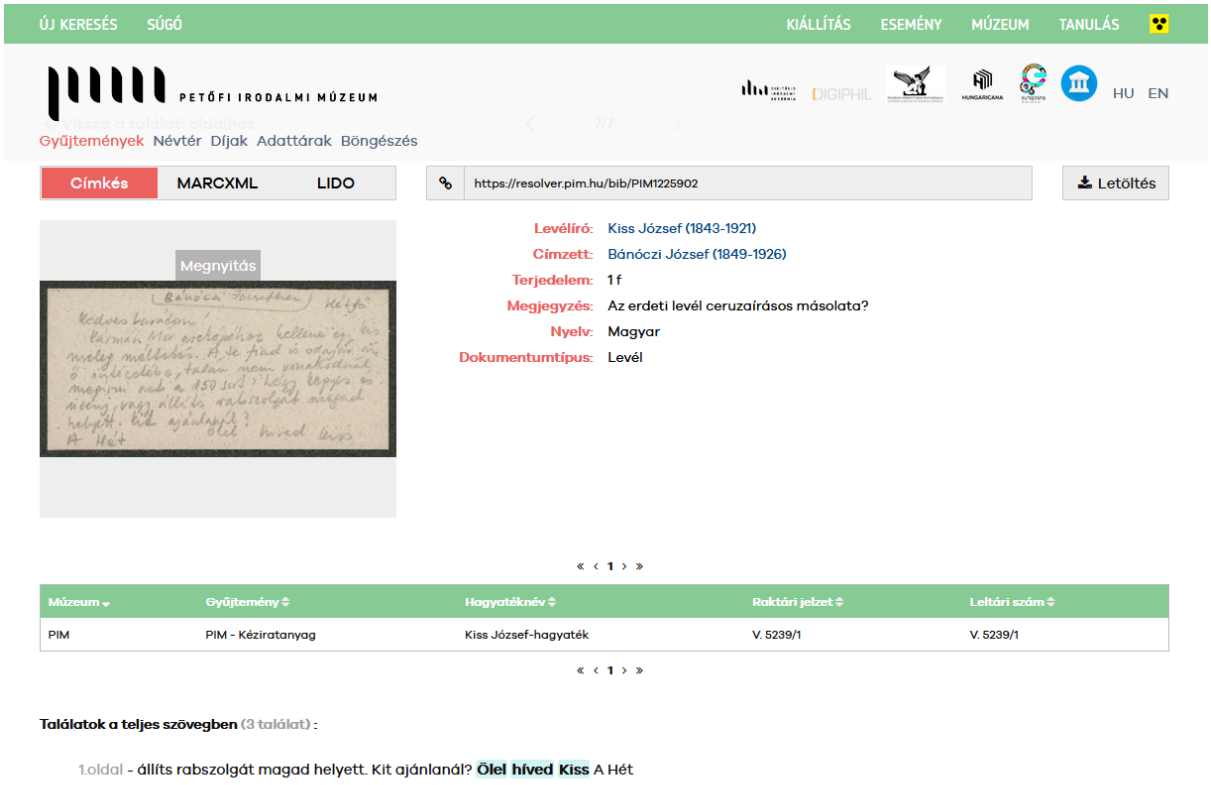
15 https://resolver.pim.hu/gyujtemeny/levelek/media/csatolt/"Kiss József 1843-1921"



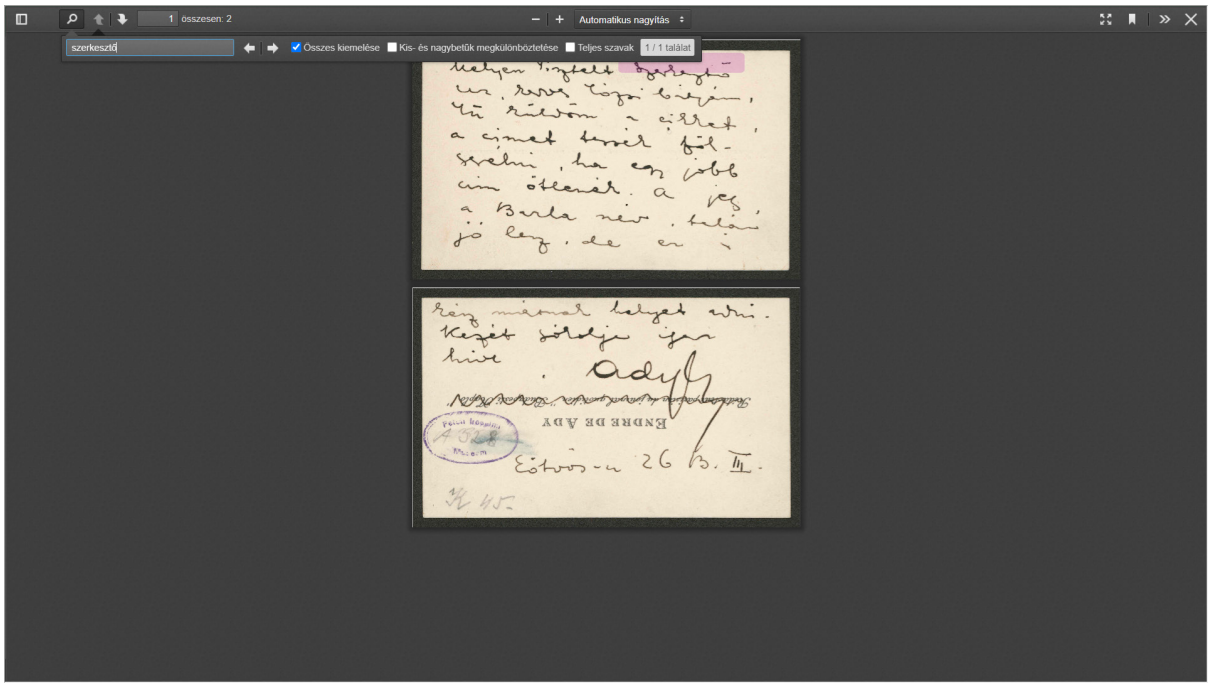
A teljes szövegű keresés esetén láthatjuk a rekordokon belüli találatokat:



Egy adott rekordra kattintva csak az azon belüli találatokat látjuk:



A PDF megnyitásakor pedig magán a kéziratfotón is ki vannak emelve a találatok:



Automatikus kézírásfelismertetés

A Transkribus program bizonyos mennyiségű kézirat manuális átírása után lehetőséget nyújt automatikus kézírásfelismertetésre használandó HTR-modellek építésére[16]. Ennek mikéntjéről a kötetben külön tanulmányunk számol be, itt csak az eredményeket villantom fel. A Kiss József kézírásából épített modell segítségével 6,94%-os hibaszázalékkal ismertettük fel automatikusan a még át nem írt leveleket. A hibák többnyire a kis- és nagybetűk, rövid-hosszú ékezetek, illetve a punktuáció hibás felismeréséből fakadtak.

Ez azt jelenti, hogy az automatikus gépi kézírásfelismertetés után minimális kézi javítás szükséges egy átirat elkészítéséhez, amely a jövőben rendkívüli mértékben felgyorsítja a szövegfeldolgozást.[17]

Publikáció 2.0 - TEI XML

Szerkesztő program

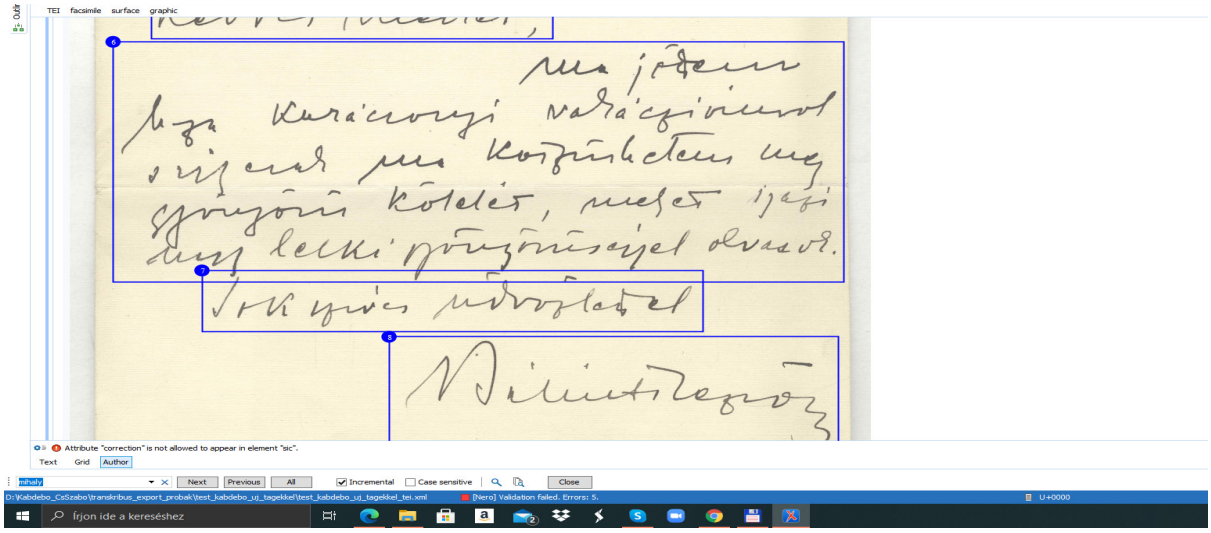
A TEI XML-ek szerkesztéséhez az Oxygen XML Editor[18] programot használjuk. Ezen belül ki lehet alakítani egy ún. frameworköt, amely az adott projekt(ek) igényeire van szabva, annak minden szempontját figyelembe véve. A szerkesztők ebben a nézetben gombok használatával tudnak annotálni, validálni, különféle automatizált műveleteket elvégezni. A framework kiterjedt textológiai-filológiai jelölésrendszer alkalmazására

16 https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/

17 A HTR modellek futtatása a Transkribus programban már fizetős funkció, amit bele kell kalkulálni a projekt költségvetésébe. Ennek fejében viszont készen kapunk egy eszközt az automatikus kézírásfelismertetés eléréséhez, a betanító szövegek előállításán túl csak a modell paraméterezését kell elvégeznünk. Az egyes modellek egymásba is építhetők, így egyre általánosabb érvényű HTR-modell hozható létre az adott nyelven.

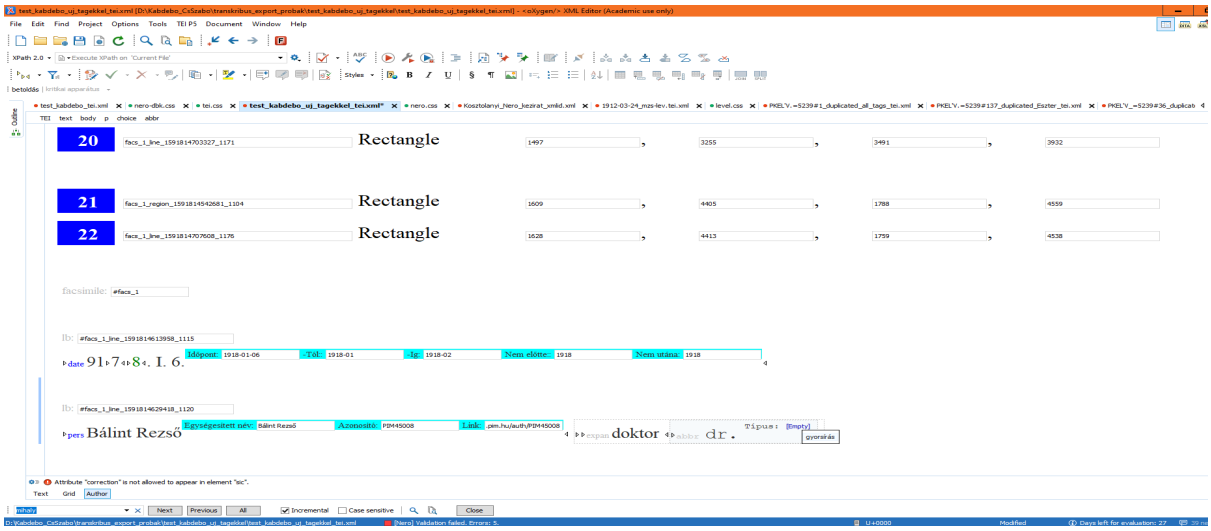
18 https://www.oxygenxml.com/xml_editor.html

nyújt lehetőséget anélkül, hogy a szerkesztőnek a technológiai hátteret ismernie kéne. Az adatgazdagítás, metaadatok részletes kitöltése, külső adatbázisok bekapcsolása, legkülönfélébb annotációk alkalmazása, a text-image linking (kép és szöveg összekötése) mind felhasználóbarát módon megoldható ebben a környezetben.



A szerkesztőfelületen a szöveg és a benne lévő annotációk megjelenítését saját igényeink szerint alakíthatjuk, akár úgy is, hogy a szerkesztő a majdani publikált formát (is) láthassa. Meghatározhatunk template-fájlokat, amelyek egy új dokumentum készítéséhez szükséges alapvető elemeket tartalmazzák, ún. sémafájlokat (DTD, RELAX NG, Schematron stb.) alkalmazhatunk, amelyek a nemzetközi szabvány mellett saját szabályrendszerünk szerint is ellenőrzik a dokumentumokat, beépíthetünk automatikus műveleteket (pl. azonosítókiosztás, metaadatok beemelése, illetve továbbítása stb.).

Az Oxygen rendelkezik ún. Git-klienssel is, amelynek segítségével a szerkesztők közvetlenül a verziókezelő rendszerbe dolgozhatnak, illetve megoldható a szerkesztőségi környezetből való közvetlen publikáció is.



Eszközök integrációja

Ezen a területen is több ponton összekötöttük a közgyűjteményi infrastruktúra egyes elemeit a szövegfeldolgozás szerkesztőségi környezetével.

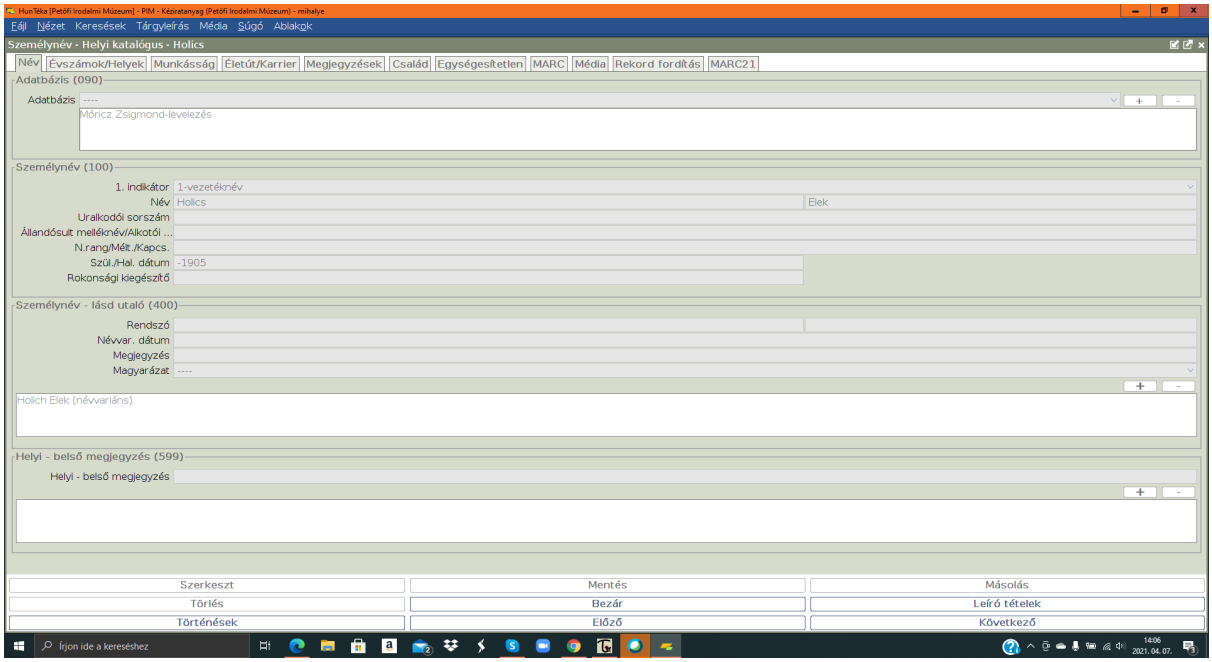
Egyfelől kialakítottuk a PIM Névtérrel való együttműködést, amelyhez a Huntéka (gyűjteménykezelő-rendszer) bizonyos fejlesztéseire volt szükség, másfelől beépítettünk az Oxygen-frameworkbe egy funkciót, amelynek segítségével a szerkesztők közvetlenül a PIM Névterében kereshetnek, illetve azon keresztül azonosíthatják a különböző entitásokat.

1. A PIM Névterével való együttműködés kialakítása első szakaszában a személynevek problematikájával foglalkoztunk, melynek során a következő kérdésekre kellett választ adnunk:

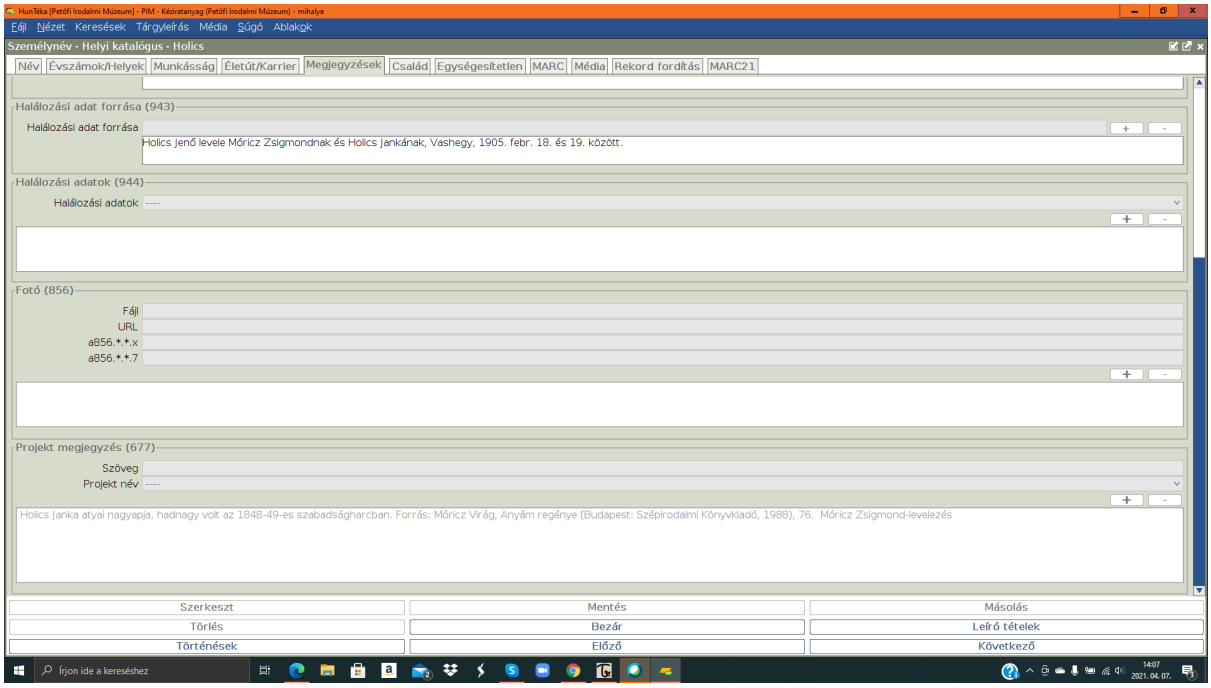
- hogyan jussanak el a projekt során előkerülő új adatok a PIM névtér szerkesztőihez?
- mi legyen azokkal a nevekkel, amelyek nem kerülhetnek be az éles névtérbe, mert még nincs elég adat, de fontosak a projektben?
- mi legyen azokkal a nevekkel, amelyek csak az adott projektben bírnak jelentőséggel?
- mi legyen azokkal a jegyzetekkel, amelyek csak az adott projekt szempontjából adekvátak?

Megoldásként kialakítottunk egy általános workflow-t a szövegfeldolgozó projektek során előállt adatok továbbításához, illetve a Huntékában a következő fejlesztéseket végeztük el:

- új 090: Adatbázis a digitális filológiai projekteknek (összes a projektben előforduló név)

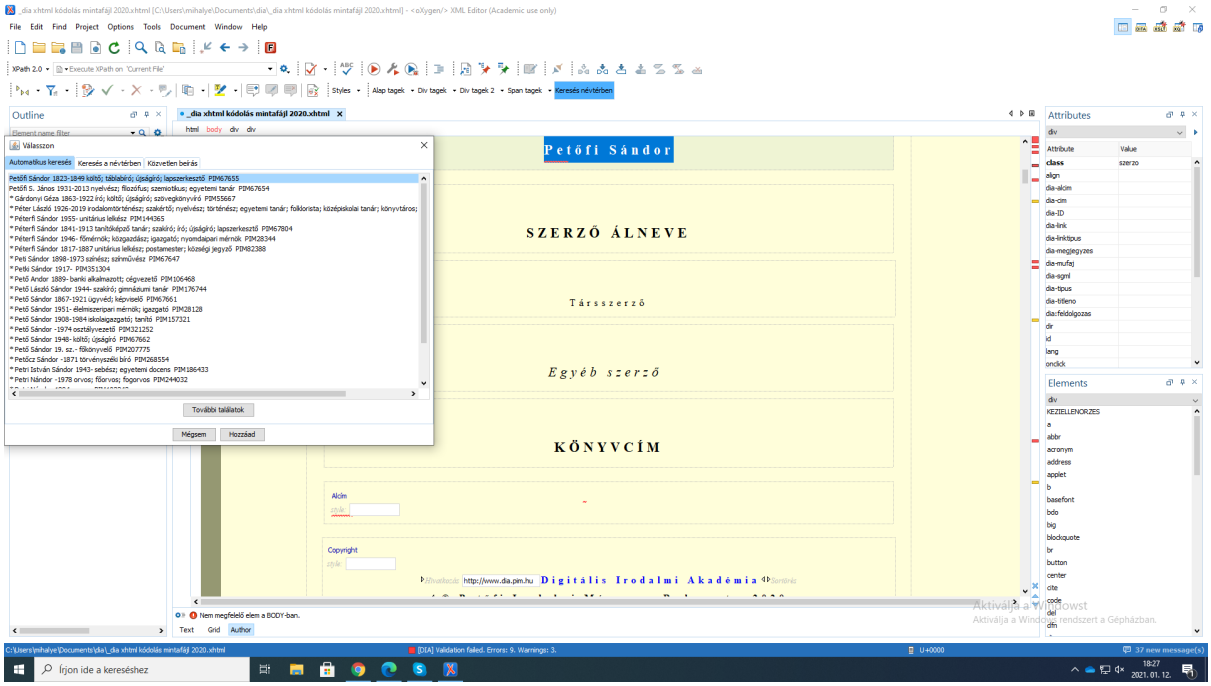


- új forrástípusok felvétele: analitikus forrás és lektorált internetes forrás
- új megjegyzés mező: projekt-megjegyzés összekapcsolva a projekttel



Ezek segítségével minden keletkezett adat helyet kapott a meglévő névtérstruktúrába. Folyamatban van a módszer kiterjesztése egyéb entitástípusokra (helynév, testületi név, terminus, műcím, bibliográfiai tétel), valamint az adatok Opac felületén való megjelenítése.

2. Az Oxygen-frameworkbe integrált névtér-funkció jelenleg a DIA (Digitális Irodalmi Akadéma[19]) frameworkjében valósult meg:



19 https://pim.hu/hu/dia

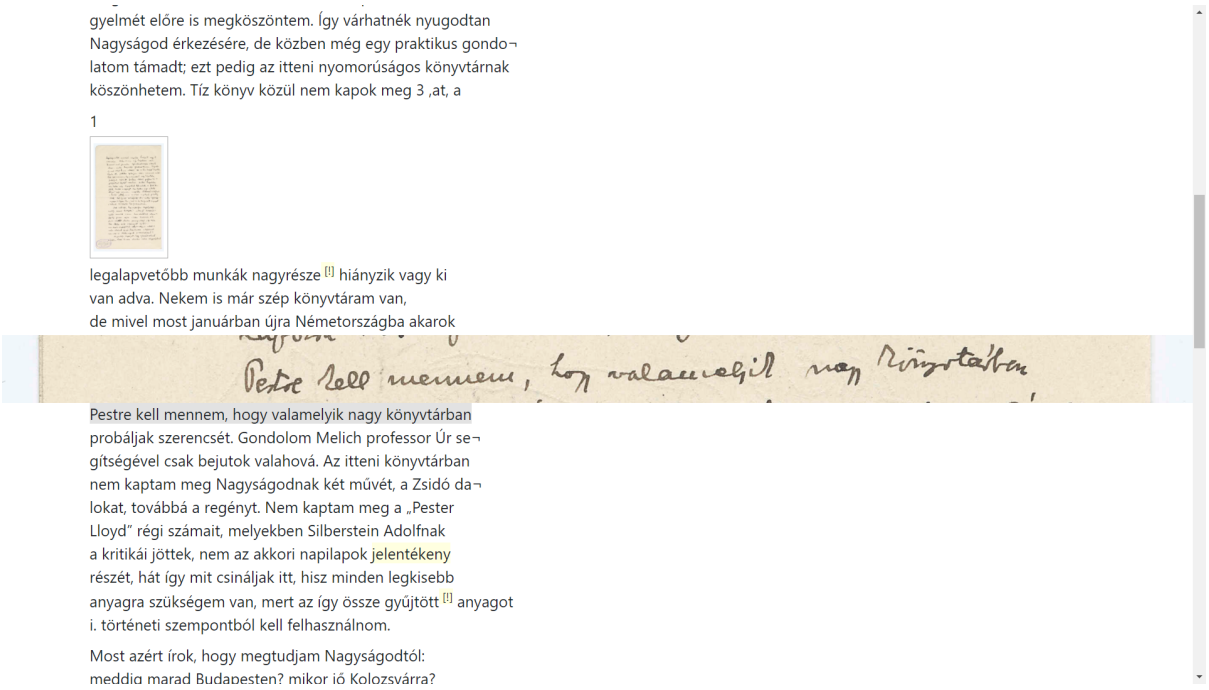
A szerkesztő a szövegben szereplő névalakot kijelölve egy kattintással beazonosíthatja a nevet a PIM Névterében, illetve egy újabb kattintással beemelheti a nevet azonosító adatokat annotációként. Folyamatban van a funkció implementálása a többi Oxygen-frameworkbe, illetve bővítése az egyéb entitástípusokra.

Publikációs, szűrési, keresési, rendezési lehetőségek

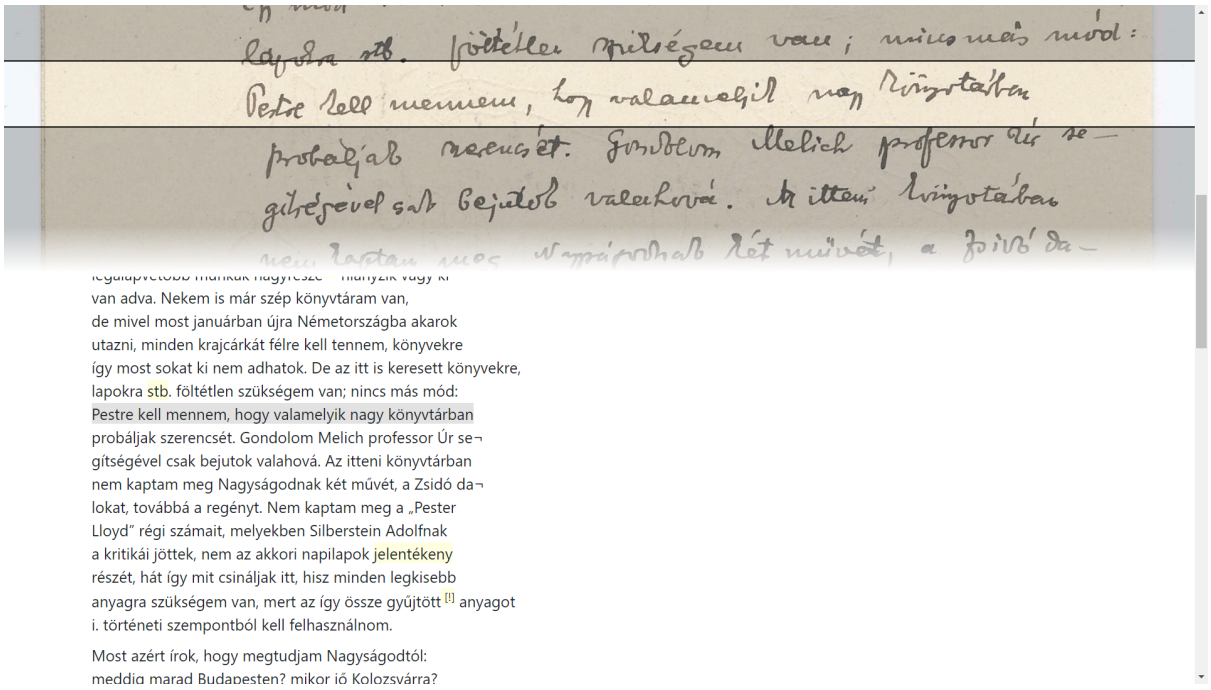
A TEI XML megjelenítési lehetőségei korlátlanok. A pilot projekt során arra koncentráltunk, hogy a felhasználók, kutatók számára leginkább hasznos vizualizációs módokat hozzunk létre. Ezért fejlesztettünk egy új eszközt a digitális fakszimile és az átírt szöveg együttes megjelenítésére, a text-image linking szövegfeldolgozó módszert elsőként alkalmazva Magyarországon. Ennek lényege, hogy a kép zónái össze vannak kapcsolva az átírt szöveggel, tehát a felhasználó együttesen láthatja a kettőt.

Háromféle nézetet is kialakítottunk:

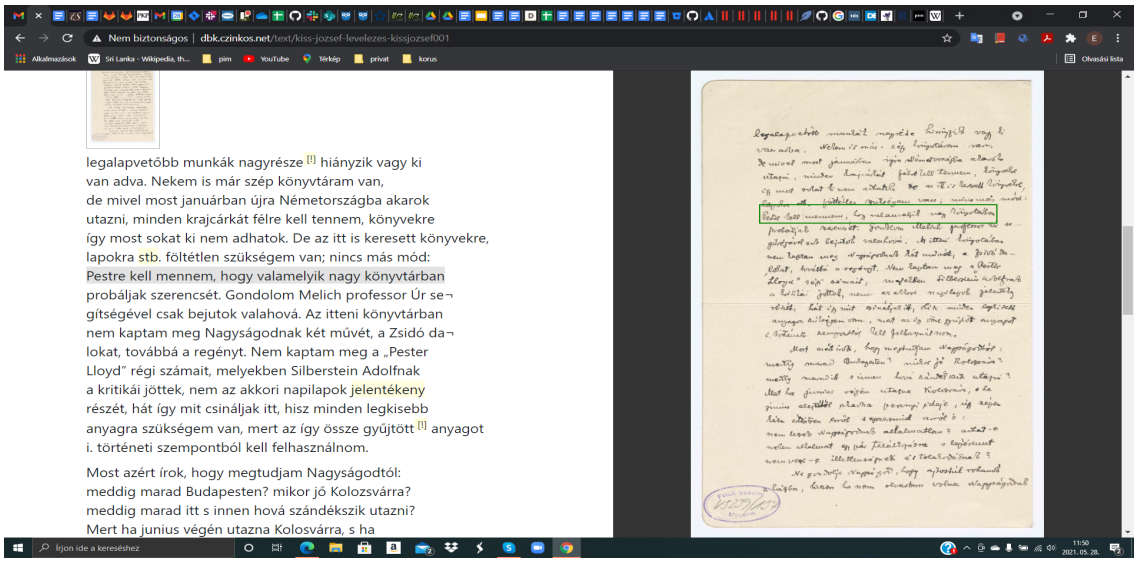
1. A kurzorral az átírt szöveg adott sora fölé menve megjelenik a fakszimile megfelelő sora:



2. A kurzorral az átírt szöveg adott sora fölé menve láthatjuk a fakszimile megfelelő részletét kisebb kontextusban:

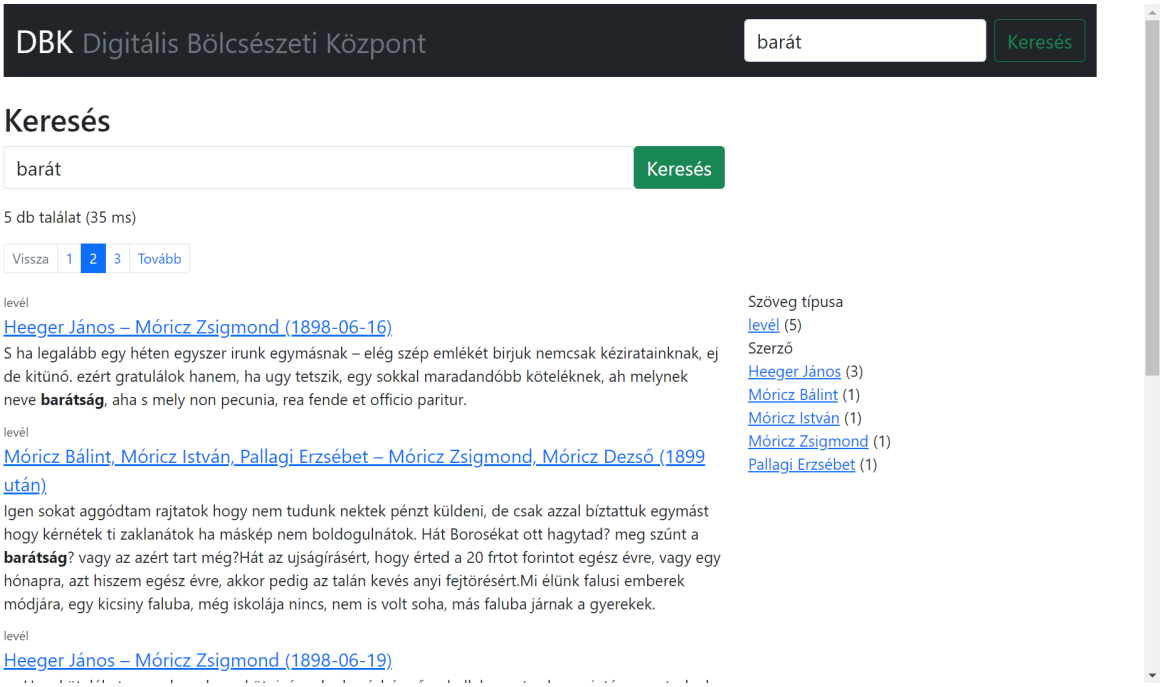


3. Osztott képernyőn láthatjuk a fakszimilét és az átírt szöveget (mindkettőn kijelölve az éppen olvasott szövegrész):



A szövegben szereplő annotációk kurzorral a (sárgával) kiemelt szövegrészek fölé menve jelennek meg.

A keresési, szűrési műveleteket itt is facetek használatával tesszük egyszerűvé, támogatóvá.



Automatikus névfelismertetés

A Nyelvtudományi Kutatóközponttal való együttműködés keretében megkezdett újabb fejlesztés eredményeképpen a szerkesztőségi környezet része lesz egy másik funkció is, amely szintén a nevek beazonosítását szolgálja NLP (Natural Language Processing[20]) eszközökkel. A Nyelvtudományi Kutatóközpont által fejlesztett e-magyar (https://e-magyar.hu/hu/) automatikus névfelismerő (Named Entitiy Recognition) moduljának[21] beépítésével lehetséges lesz a szövegekben szereplő nevek automatikus megjelölése, méghozzá névtípusonként. Ezután következhet a fentiekben felvázolt beazonosítási procedúra.

Amennyiben nincs kapacitás a nevek manuális beazonosítására, a technológia mindenképp alkalmazható a publikációs felület keresőjébe beépítve, így a felhasználó a program által névként felismert szóalakok között tud böngészni.

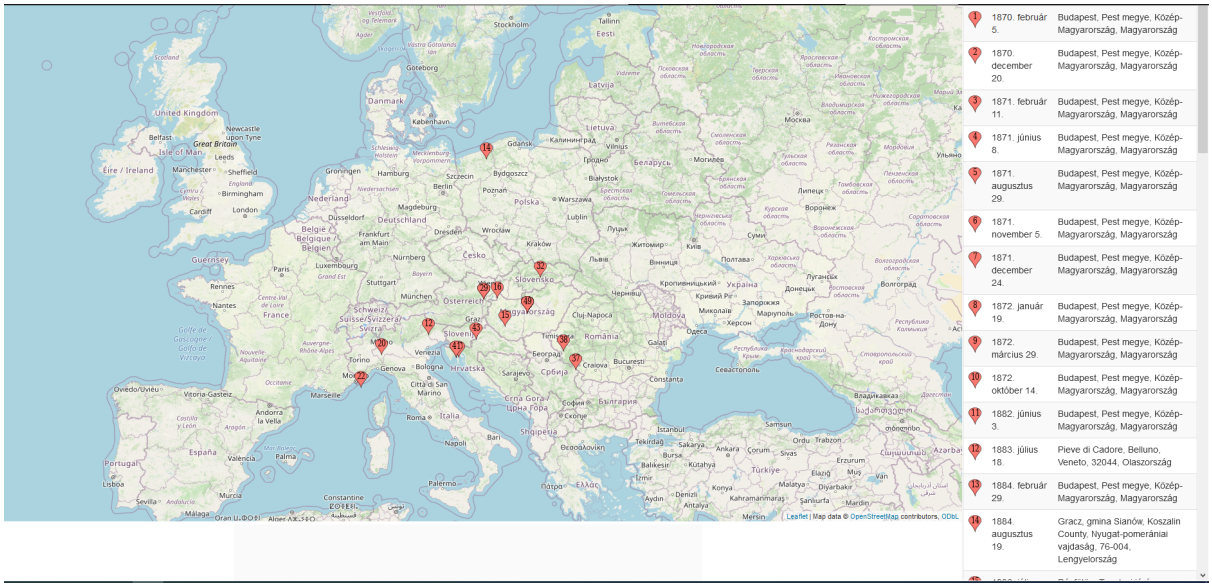
Adatvizualizációs lehetőségek

A metaadatok, illetve a szövegekben elhelyezett annotációk felhasználásával a legkülönbözőbb statisztikák, vizualizációk állíthatóak elő az egyes szövegkorpuszokról.

Levelezéskiadás esetén kézenfekvő a levelek feladási, illetve fogadási helyszíneinek térképen való nyomon követése:

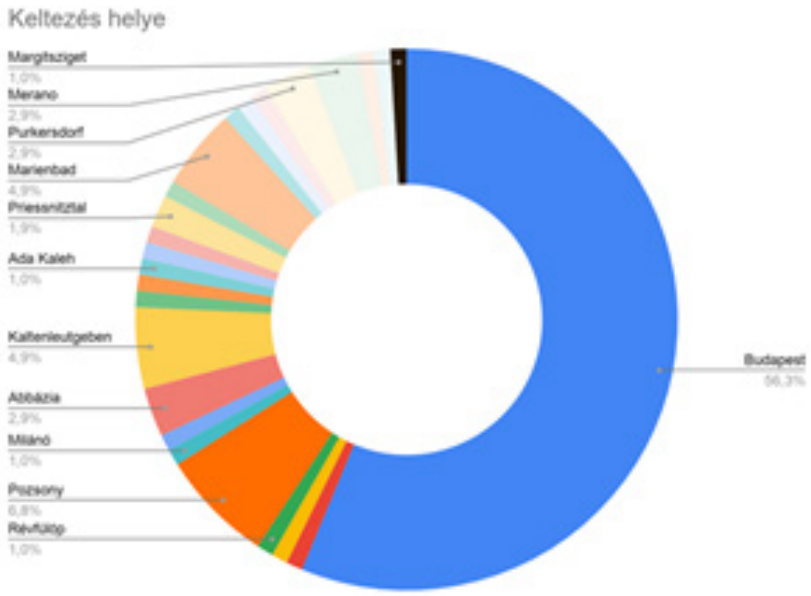
20 https://en.wikipedia.org/wiki/Natural_language_processing

21 https://e-magyar.hu/en/textmodules/emner



Kiss József levélkeltezései. Ahol Kiss József élete során megfordult.

Érdekelhet minket a levélkeltezések helyszíneinek aránya:



Vagy kimutatást készíthetünk a levelezőpartnerek nemek szerinti eloszlásáról is:

Kiss József levelezőpartnerei:
Zöld: férfi
Sárga: nő

A lehetőségek száma végtelen, csak a képzeletünk szabhat határokat. Fontos azonban mindig szem előtt tartani, hogy a vizualizációk az adatok olyan aspektusait világítsák meg, amelyek anélkül nem, vagy nagyon nehezen értelmezhetőek. Hiszen a végső cél minden esetben a felhasználók számára hasznos szolgáltatások nyújtása.