**Cséve Anna**

*Petőfi Irodalmi Múzeum*

cseve.anna@pim.hu

**Fellegi Zsófia**

*Petőfi Irodalmi Múzeum*

fellegi.zsofia@pim.hu

**Kómár Éva**

*Magyar Nemzeti Múzeum*

komar.eva@mnm.hu

# Móricz Zsigmond levelezésének (1892–1913) digitális kritikai kiadása: Esettanulmány

2016-ban indult el a Petőfi Irodalmi Múzeumban az a hároméves NKFIH-projekt, amely Móricz Zsigmond levelezésének (1892–1913) digitális kritikai kiadását tűzte ki célul. A feladat kihívást jelentett a Móricz-műhely számára, hiszen a korábbi, papíralapú kiadási gyakorlatra csak részben támaszkodhattak. A múzeumi informatikai lehetőségek, a filológiai problémák és az alkalmazott szoftverek párbeszédéről szóló esettanulmány a projekt első évének problémafelvetéseiről, megoldásairól szól. Nem törekszik teljes áttekintésre, hiszen munkafolyamat közben ad hírt egy formálódó gyakorlatról.

Kulcsszavak:
digitális filológia, Móricz Zsigmond, kritikai kiadás, levelezés, DigiPhil

## 1. Bevezetés

A Petőfi Irodalmi Múzeumban (PIM) a tudományos igényű szerzői szövegkiadásoknak hagyománya van, a múzeum többek között Móricz Zsigmond életművének számos forrását, naplóit jelentette meg az elmúlt évtizedben. A modern magyar irodalom más klasszikusainak jelenleg folyamatban lévő kutatásait, posztumusz szövegkiadását tekintve Móricz Zsigmond levelezésének szisztematikus filológiai feltárása is régóta elvégzendő feladat. Ennek érdekében fontos előrelépés történt 2016-ban, amikor megkezdődött a Móricz-levelezés kritikai kiadása az NKFIH támogatásának köszönhetően.[1] A levelezés nagyságrendjét jelzi, hogy a hagyatékban (a PIM különgyűjteményében) található, Móricznak címzett levelek, illetve a rokonoknak írt vagy másolatban

[1] Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal (NKFIH), szerződés nyilvántartási száma: 116201.

fennmaradt szerzői levelek száma 9540 darab. A kiadásnak ugyanakkor eleve számolnia kell a levelezéskorpusz önfeltáró jellegével, s a forrásfeltárás a szövegleírással párhuzamosan ma is folyik. A pályázat beadásával egy időben zajlott a PIM-ben a digitális kritikai kiadásnak mint szolgáltatásnak a továbbfejlesztése, így az NKFIH-projekt már beadásakor tartalmazta a DigiPhil digitális szövegkiadási műhely tapasztalatait, integrálta fejlesztési ambícióit. A digitális közzététel a Móricz-kutatásban kísérleti módszernek számít.[2]

A projekt során olyan új típusú feladatokat kellett megoldani, melyekhez a kutatócsoport hagyományos és digitális filológiai tudását egyesíteni kellett. A feladat a digitális műhely részéről is kihívásnak számított, hiszen a DigiPhil addig csak könyvalapú kritikai kiadások online közzétételének specifikációjával foglalkozott. A szoftverkörnyezet a DigiPhil-ben már alkalmazott komponensekre épül, a feldolgozás módszereit ehhez igazítva alakítottuk ki. Móricz Zsigmond leveleinek digitális kiadása több fázisú munkafolyamatként indult el: fő kérdése az volt, hogyan tudja megteremteni saját munkafolyamatának és a digitális kritikai kiadásnak egymással összekapcsolódó komplex informatikai hátterét. A múzeumi informatikai lehetőségek, a tisztán filológiai problémák és az alkalmazott szoftverek párbeszédéről szóló esettanulmány a projekt első évének problémafelvetéseiről, megoldásairól szól, munkafolyamat közben ad hírt egy formálódó gyakorlatról.

## 2. A projekt megvalósulásának lépései

### 2.1. A levelek digitalizálása

A Móricz-levelezéskiadás kéziratkatalógus hiányában a *Huntéka-M* könyvtári-múzeumi integrált rendszerre[3] támaszkodik, amely a hasonló projektek esetében még nem gyakori eljárás. A PIM gyűjteménykezelő rendszere[4] tartalmazza a PIM-ben található kéziratok strukturált alapadatait (levélíró, címzett, dátum, terjedelem, állapotleírás, nyelv). A 2676 szerzői és 6864 írónak címzett levél metaadatainak exportálása után táblázatos listák készültek, így egy felületen vált megjeleníthetővé a levelezés időrendje.

Első lépésben a hároméves projekt során kiadni kívánt, 1923-ig terjedő időszak leveleinek másolatait állítottuk elő e lista alapján: Móricz Zsigmondnak szóló 1582 levelet, Móricz 314 levelét és 788 keltezetlen levelet. A válogatás alapján 2684 kézirat digitalizálását végeztük el. Móricz Zsigmond leveleinek 93 százaléka még kézírással készült, ugyanez elmondható a Móricznak címzett levelekről is. A fakszimilék jelentős segítséget adtak az átiratok elkészítéséhez, megfeleltek az archiválási és állományvédelmi szempontoknak, hiszen az eredeti dokumentumokat nem kellett újra átmozgatni az ellenőrzési munkafolyamat során. 2017-ben magángyűjteményekben is folytattuk a feltárásokat, több száz ismeretlen levélről a helyszínen készítettünk digitális másolatokat. A kutatás során feltárt új források a feldolgozandó korpusz darabszámát jelen-

[2] DigiPhil: A magyar irodalomtudomány filológiai portálja, hozzáférés: 2018.04.12, `http://digiphil.hu/`.

[3] *Múzeumi Huntéka*, hozzáférés: 2018.04.12, `https://qulto.eu/muzeumi-hunteka`.

[4] Petőfi Irodalmi Múzeum Huntéka-M online felülete, hozzáférés: 2018.04.12, `https://opac.pim.hu/`.

 DOI 10.31400/dh-hun.2018.1.227

tősen megnövelték, ezért a kutatásütemezést át kellett dolgozni. A Móricz-levelezés kritikai kiadása az 1913-as évvel zárul, ebből az évből 1 225 dokumentum ismert.

A mesteranyag tárolási formátuma a TIFF (Tag Image File Format), mivel veszteségmentesen tömörít, és alkalmas a képek metaadatainak tárolására is. A mesterfájlok biztonságos elhelyezéséről a PIM informatikai rendszere gondoskodik. A levelek digitális fakszimiléi a szolgáltatás keretein belül megtekinthetők lesznek, a digitális kiadás szövege mellett található ikonra kattintva az eredeti forrás digitális másolata összevethetővé válik az átirattal. A kéziratok darabszámához képest sokkal több, közel 7000 képfájl készült: a képeslapok és borítékok mindkét oldaláról, a levelek üres oldalairól is.

### 2.2. Szövegkódolás

A filológiai szempontokat tekintve legfontosabb feladat a metaadatok leírása és a szövegjellemzők azonosítása, a céloknak megfelelő jelölőnyelv alkalmazása. A Móricz-kutatócsoport a TEI (Text Encoding Initiative) XML (Extensible Markup Language) ajánlása[5] mellett döntött, alkalmazkodva ezzel a DigiPhil korábbi gyakorlatához, amely megfelel a nemzetközi elvárásoknak. Elsődleges szempont volt a szövegekben előforduló sajátosságok szofisztikált jelölése, a kritikai kiadás filológiai alapvetése szerint. A projekt TEI elemkészletét a DigiPhil alakította ki a kutatócsoport igényeinek megfelelően, a TEI P5 kéziratleírásra kidolgozott modulja szerint (Manuscript Description).[6]

A levelezéskiadások általában csak a levél szövegét és annak hordozóját írják le, nem foglalkoznak például a borítékkal vagy a mellékletekkel, pedig a mellékelt rajzok vagy versek relevánsak a korpusz egészének szempontjából. A boríték, a levél és a melléklet egy objektum egyes részeinek tekinthetők, leírásuk ennek a sorrendnek megfelelően történik. A kutatócsoport nemzetközi digitális levelezéskiadások gyakorlatában nem talált példát ilyen típusú leírásra, ugyanakkor meg tudtuk oldani, hogy a TEI-fájlban ez a három egység a metaadatok és a szövegkódolás szintjén egyaránt reprezentálható legyen.

Moduláris felépítése miatt a TEI kódolásában mindez egy fájlon belül is megoldható a metaadatok szintjén:

```
<msPart style="envelope">
<msPart style="letter">
<msPart style="attachment">
```

Hasonlóan a szövegleírás szintjén is:

```
<div type="envelope" style="handwritten">
<div type="letter" style="handwritten">
<div type="attachment" style="handwritten">
```

[5] „TEI: P5 Guidelines," Text Encoding Initiative, hozzáférés: 2018.04.12, http://www.tei-c.org/Guidelines/P5/.

[6] „Manuscript Description," P5: Guidelines for Electronic Text Encoding and Interchange 3.3.0, 2018. jan. 31., hozzáférés: 2018.04.12, http://www.tei-c.org/release/doc/tei-p5-doc/en/html/MS.html.

A fenti hármas struktúrán túl a kéziratszöveg átírásának linearitása további problémákat vetett fel. Például idegen szöveg beékelődése esetén az időrend volna mérvadó, ám ez sokszor nem rekonstruálható egyértelműen. Ilyen esetben, vagyis, ha a szövegszegmentumok sorrendje vitatható, a sorrend meghatározásakor a levél struktúráját követjük, olyan szempontokat is figyelembe véve, mint az olvasás iránya.

A levéltől részben elkülönülő paratextuális szövegszegmentumok, mint a postabélyegző vagy a pecsét, alapvető információkat tartalmazhatnak. A kritikai kiadás objektumstruktúrája így a következőképpen alakult: boríték, levél (ezen belül: fejléc, nyomtatvány, rajz, lábléc, pecsét), melléklet. A kéziraton található szövegszegmentumok elhelyezkedését nem szükséges jelölni a fenti struktúrán kívül, ezekről ugyanis a kéziratokról készült fakszimile nyújt információt.

A jelölőnyelvi leírás számos szövegjellemzőt rögzít (pl. betoldás, aláhúzás, javítás, idegenkezűség). Az elírásokat, betűkimaradásokat és -tévesztéseket, az értelemzavaró helyesírást, a nehezen értelmezhető rövidítések feloldását, régi szavak, szóalakok rövidítéseit a könnyebb olvashatóság érdekében a betű szerinti átirat megtartása mellett emendáltuk. Az online kiadásban a betűhű átirat és az olvasószöveg egyaránt olvasható lesz.

A levélküldés folyamata előtti (pl. használt papírra írt) és utáni rájegyzéseket nem a főszövegben, hanem a szerkesztői jegyzetben tesszük közzé: a kézirattest elő- vagy utóélete részének tekintjük, s mint a levél formai jellemzőjét a levélszöveg metaadatai között szereplő levélleírásban szerepeltetjük. Jellegzetes példa Pallagi Gyula Móriczcznak szóló 1900. november 18-a után keletkezett levelén Móricz *A szép lány suttog...* kezdetű, mindeddig ismeretlen *Pua* címmel emlegetett versének két versszaka.[7]

Ahogy más projektek esetében láthattuk, az egyes kiadások a TEI-ajánlásokat alapul véve saját sémákat hoznak létre.[8] A Móricz-levelezés kódkészleténél is több egyedi megoldás született. Ezek közül egyetlen példát szeretnénk kiemelni, amelynek bevezetése a levélkéziratokon található szövegszegmentumok bonyolult felépítésének köszönhető.

A TEI logikája az egyes szövegszegmentumok elkülönítésére több megoldást tesz lehetővé. A Móricz-kiadásban például az idegenkezűség jelölése eltér a nemzetközi gyakorlattól: idegenkezűség esetén általában a `<handShift/>`, illetve az `<anchor/>` jelölőt alkalmazzák. Ezek a jelölők azonban pontszerűek, így ezeknek a szegmentumoknak a kiemelése és vizualizálása informatikai szempontból komoly nehézséget okozott volna. A `<seg>` jelölő használatával ezt könnyedén elkerülte a kutatócsoport, anélkül, hogy megsértette volna a TEI ajánlását. A `@corresp` attribútum segítségével lehet megadni a levélre rájegyző nevét, így ezek az információk összekapcsolódnak és kereshetővé válnak, a `@type` attribútumban pedig a főszövegben található szegmentum jellegét (pl. titkosírás, pecsét) lehet definiálni.

```
<p><seg type="handShift" corresp="Móricz Zsigmond"></seg></p>
```

[7] Pallagi Gyula levele Móricz Zsigmondnak, Budapest, 1900. november 18. után, PIM Kézirattár, M. 130.

[8] A Vincent Van Gogh-levelezés kiadás készítésekor például kiegészítették a TEI-sémát saját jelölőkkel. Leo Jansen, Hans Luijten and Nienke Bakker, eds., *Vincent van Gogh – The Letters.* Version: December 2010. Amsterdam & The Hague: Van Gogh Museum & Huygens ING., `http://vangoghletters.org/vg/about_6.html`.

 DOI 10.31400/dh-hun.2018.1.227

A <seg> elem nemcsak az idegenkezűség jelölésére szolgál; így jelölendő az aláírás (<seg type="signature">Dr László</seg>) és a pecsét (<seg type="stamp">).

Szintén a TEI-ajánlástól eltérő megoldás a levél zárlatának jelölése. A TEI a <closer> címkét javasolja, azonban a szintaxis alapján ezt más szövegrész nem követhetné. A problémát a kutatócsoport úgy oldotta meg a <closer> kihagyásával, hogy az aláírást követő szövegrészeket, mint például a lábléc (<floatingText type="footer">), a TEI által szorosan nem definiált szövegszegmentumok leírására szolgáló címkével jelölte.

### 2.3. A TEI-fejléc lehetőségei

A kéziratra vonatkozó adatokból a projekt szempontjából releváns metaadatok körét a kutatócsoport állapította meg. A kritikai kiadás jellegéből fakadóan a formai leíráshoz használt mezőkészlet jóval gazdagabb, mint az a könyvtári bibliográfiai feldolgozásnál megszokott. Így például három külön adatelem a megírás helye, a feladás helye és az átvétel helye, de ugyanez érvényes a dátumot leíró mezőkre is, a kutatócsoport megadja a megírás, a feladás és az átvétel dátumát. A múzeum könyvtári adatbázisának *kézirat* űrlapján a hely és dátum leírásához csak a *keletkezés helye* és *ideje* HUNMARC-mezők (ismételhető c260$a és c260$c) állnak rendelkezésre.

A TEI-ajánlás szerint az XML-fejléc része részletes metaadat-rögzítésre ad lehetőséget, sőt lehetőséget nyújt adatgazdagításra is. Az információk hozzáadása a metaadatokhoz egyrészt saját erőforrásból, a háttéradatbázisok segítségével történik, másrészt külső tezauruszok, névterek bekapcsolásával. Az adatgazdagítás négy fő dimenziója (személy, hely, idő, fogalom) közül jelenleg a személy, a hely és a fogalom vonatkozásában történik bővítés.

A TEI-fejléc nyitó és záró címkéje közötti rész tartalmazza az objektum metaadatait. A <teiHeader> alatt a <fileDesc> foglalja össze a digitális kiadásra vonatkozó információkat. A <title> elemen belül a kiadás címét:

```
<titleStmt>
 <title>Móricz Zsigmond levelezés kritikai kiadás</title>
</titleStmt>
```

A <publicationStmt> rész jól reprezentálja a finomítási lehetőségeket. A neveket jelölő elemeken belül megadható, hogy személyről, intézményről vagy helységnévről van szó (<persName>, <orgName>, <placeName>).

Itt már látható példa az adatgazdagításra is URI-k megadásával a <ref> elem @type attribútumában: a kiadó nevéhez bekerült a PIM VIAF[9] katalógusában lévő azonosítója, valamint a kiadás helyénél a GeoNames[10] egyedi azonosítója. Hasonlóképpen hivatkozik az <availability> címke a közzétételi jogokra.

Nagyon fontos megadni a feldolgozott objektum perzisztens egyedi azonosítóját (PID) és URI-ját (Uniform Resource Identifier), ugyanis ezek az azonosítók garantálják az egyes digitális objektumok (jelen esetben a levelek) azonosíthatóságát és a kiadás idézhetőségét. Ezeket szintén a @type attribútum jelöli az <idno> elemben.

[9] Virtual International Authority File, hozzáférés: 2018.04.12, http://viaf.org/.
[10] GeoNames, hozzáférés: 2018.04.12, http://www.geonames.org/.

```
<publicationStmt>
 <publisher>
  <orgName>Petőfi Irodalmi Múzeum</orgName>
  <ref type="url">http://viaf.org/viaf/152132060/</ref>
  <ref type="url">http://www.pim.hu</ref>
 </publisher>
 <pubPlace>Budapest <ref type="url">http://www.geonames.org/
  3054643</ref>
 </pubPlace>
 <date>2015</date>
 <availability>
  <p>©Free Access - no-reuse <ref type="url">http://www.europeana.eu/
rights/rr-f/</ref>
  </p>
 </availability>
 <idno type="PID">o:PKEL.M.100-2553-18_a</idno>
 <idno type="URL">o:PKEL.M.100-2553-18_a</idno>
</publicationStmt>
```

A dokumentum egészének leírása a <sourceDesc> elem alatt található további strukturált egységekben. Az <msDesc> (manuscript description) címke jelöli a kézirat metaadatainak leírására vonatkozó információkat. A lelőhely megadásánál az <msIdentifier> elemnél lehetne leírni a provenienciára vonatkozó információkat, de mivel a közgyűjteményekben ezek érzékeny adatoknak számítanak, a TEI-kódban és a DigiPhil oldalán az adatok nem lesznek nyilvánosak, egyelőre csak a hagyaték neve szerepel az <msName> alatt felvéve.

```
<sourceDesc>
 <msDesc>
  <msIdentifier>
  <country>Magyarország</country>
  <settlement>Budapest<idno type="KOHA_GEO">KOHA_GEO:9227</idno>
  </settlement>
  <institution>Petőfi Irodalmi Múzeum</institution>
  <repository>Petőfi Irodalmi Múzeum Kézirattár</repository>
  <idno>PIM M. 100/2553/18</idno>
  <msName> Móricz Zsigmond-hagyaték </msName>
 </msIdentifier>
```

Az objektum egyes elemeit az <msPart> elem szegmentálja. A levél fizikai leírását a <physDesc> címke vezeti be, melynek további részeivel megadható a levél mérete és állapota.

```
<msPart style="letter">
 <msIdentifier/>
  <physDesc>
```

 DOI 10.31400/dh-hun.2018.1.227

```
   <objectDesc>
    <supportDesc>
     <extent>
      <measure type="quantity" unit="folio"> 4 </measure>
      <dimensions unit="mm">
       <height> 109 </height>
       <width> 174 </width>
      </dimensions>
     </extent>
     <condition>
      <p>Sárgult papíron.</p>
     </condition>
    </supportDesc>
   </objectDesc>
  </physDesc>
</msPart>
```

A levél részletes leírása a <profileDesc> címke alatt látható. Ebbe a részbe került adatgazdagítás céljából egy formai tárgyszó a Getty Art & Architecture (AAT)[11] tezauruszából. A TEI-fejlécnek ebben a részében található a megírásra, a feladásra és az átvételre vonatkozó metaadatok feltüntetése a <creation>, <correspAction type="sent"> és a <correspAction type="recieved"> jelölők segítségével.

```
<profileDesc>
 <langUsage>
  <language ident="hu"/>
 </langUsage>
 <textClass>
  <keywords scheme="AAT" corresp="Letter">
   <term>levél</term>
   <idno type="AAT"> AAT:300026879 </idno>
  </keywords>
 </textClass>
 <creation>
  <date when="1905-07-02"/>
  <placeName>Budapest <idno type="KOHA_GEO">KOHA_GEO:9227</idno>
  </placeName>
 </creation>
 <correspDesc>
  <correspAction type="sent">
   <persName>Móricz Zsigmond <idno type="KOHA_AUTH">KOHA_AUTH:120256
  </idno>
   </persName>
```

[11] „Art & Architecture Thesaurus," The Getty Research Institute, hozzáférés: 2018.04.12, http://www.getty.edu/research/tools/vocabularies/aat/.

```
  </correspAction>
  <correspAction type="recieved">
   <persName>Holics Janka <idno type="PIM">PIM:1153120</idno>
   </persName>
  </correspAction>
 </correspDesc>
</profileDesc>
```

### 2.4. Névterek

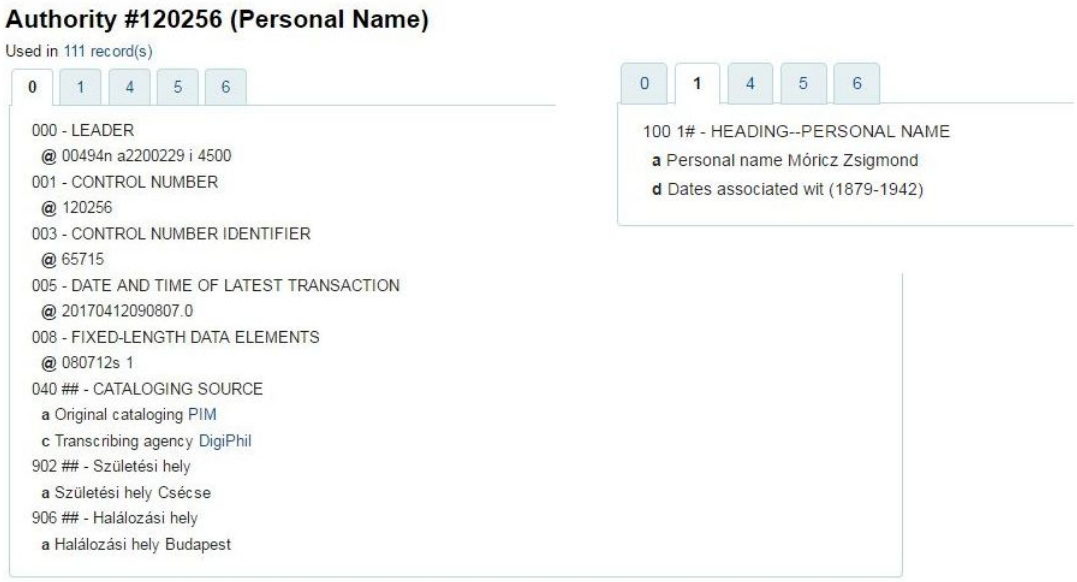
Az egységes besorolási adatok (személy- és helynevek) a szolgáltatás hozzáférési pontjait biztosítják és a szemantikai kapcsolatok kiépítését segítik. A Móricz-levelek átírásakor jelenleg a személyek, a földrajzi helyek és a műcímek azonosítását végzik a munkatársak. Az identifikáció része, hogy az entitások egyedi, állandó azonosítót kapnak, így az összes előfordulásuk kereshetővé válik.

```
090 __ a INT
a IPA
a ITO
a IKN
a TLA
a GYN
100 1_ a Móricz
j Zsigmond
d 1879-1942
400 __ Q adateltérés[$j]:Zsigmond (?)
400 __ a Zsiga
j bácsi
400 __ a M.
j Zs.
500 1_ a Holics
d 1883-1925
j Janka
667 __ a író
667 __ a újságíró
667 __ a publicista
667 __ a lapszerkesztő
680 __ a sírhely: 34. parcella
900 __ a 1879. VI. 29.
902 __ 3 ITE-9395
a Csécse
```

**1. ábra**. Móricz Zsigmond besorolási rekordjának HUNMARC mezői a Huntékában
Forrás: *Huntéka-M*, PIM

 DOI 10.31400/dh-hun.2018.1.227

A PIM személynévtér-állománya hozzávetőlegesen 600000 rekord.[12] Az életrajzi típusú adatbázisok közül a *Magyar Életrajzi Index*[13] rendelkezik érvényes, kontrollált személynév-rekordokkal, ezért a projekt számára a Huntéka-rendszerből ezt a részt migrálták a DigiPhil mögött működő könyvtári rendszerbe, a *Koha*ba.[14] A PIM névterében lévő névrekordok sok hozzáadott információt tartalmaznak (rokoni kapcsolatok, lakhelyek, temetés helye stb.), így a megfeleltetésnél redukálni kellett a kiegészítő adatok körét. Csak annyi mező került át a *Koha*-névrekordokba, amennyi minimálisan elégséges ahhoz, hogy egy személy azonosítható legyen: vezetéknév, keresztnév, a születés és halálozás adatai.




**2. ábra**. Móricz Zsigmond besorolási rekordja a *Koha*ban. Forrás: `http://biblio--intra.digiphil.hu/`

A Móricz-projekt a PIM és a *Koha* azonosítóit használja a személynevek egyértelműsítésére.

```
<persName>Édesapám<idno type="PIM" corresp="Móricz Bálint">PIM:
   297674</idno></persName>
```

Vannak olyan esetek azonban, amikor nincs elegendő adat a személy azonosításához, és így nem lehet érvényes névrekordot létrehozni az adatbázisban. A levelek irodalmi, művészeti, közéleti kapcsolatokat felvonultató adatai mellett családi vagy személyes levélváltásokra is nagy mennyiségben van példa a Móricz-levelezésben. A rokonok, barátok leveleiben sokszor előfordul csak keresztnévvel említett személy, például a

[12] Bánki Zsolt, Mészáros Tibor, Németh Márton és Simon András, „Azonos személyekre vonatkozó név besorolási rekordok automatikus felderítése a PIM adatbázisában," *Tudományos és Műszaki Tájékoztatás* 63, 12. sz. (2016): 471.

[13] *Magyar Életrajzi Index*, hozzáférés: 2018.04.12, `https://opac-nevter.pim.hu/`.

[14] *Koha Library Software*, hozzáférés: 2018.04.12, `https://koha-community.org/`.

Móricz-háztartásban segédkező alkalmazottak (Anna cseléd) vagy Móricz testvérének osztálytársai (pl. Sanyi). Ezek az entitások nem kerülnek be az adatbázis besorolási állományába, de a kritikai kiadásban fontos a megkülönböztetésük, ezért ún. lokális azonosítót (LOK) kapnak.

```
<persName>Anna<idno type="LOK" corresp="Anna_cseléd">LOK:00013
</idno></persName>
```

A helynevek és a műcímek identifikációja hasonló módon történik. A helynevek azonosításához a Geotaurusz[15] rekordjait importáltuk a *Koha*ba, így a levelek szövegeiben lévő helynevek egyedi azonosítói a *Koha*ból kerülnek a TEI-be.

```
<placeName>Gödöllőig<idno type="KOHA_GEO"
corresp="Gödöllő">KOHA_GEO:21799</idno></placeName>
```

## 3. Tervek

### 3.1. Kommunikáció integrált rendszerekkel *Huntéka-M, Koha*

A kéziratok feldolgozásának alapja a levelek formai feltárásának elvégzése, és a szövegekben előforduló entitások (a személyek, a földrajzi helyek és a címek) azonosítása.

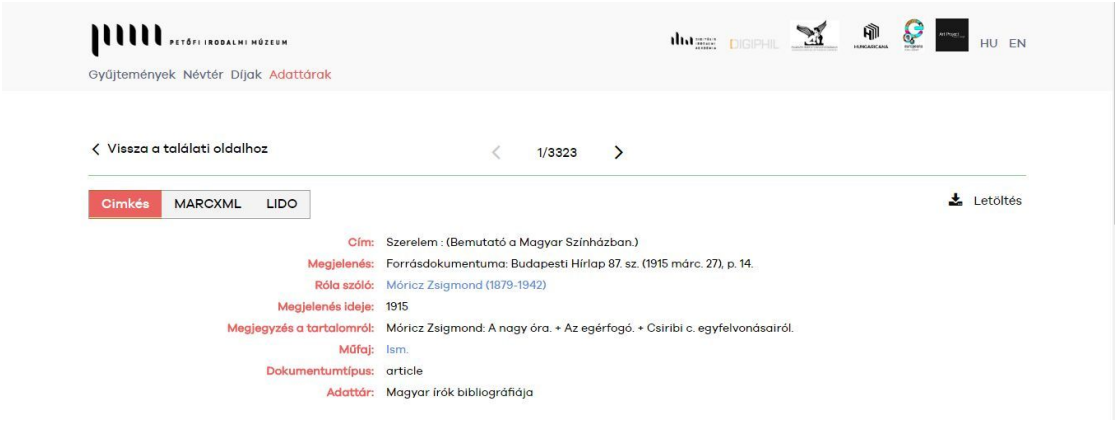
A *Huntéka-M* rendszerében nemcsak a gyűjtemények anyaga található, hanem a PIM tevékenységéhez tartozó háttérkutatások eredményei is. A múzeum jelentős személynévtérrel rendelkezik, de emellett egyéb bibliográfiai és faktográfiai jellegű adatbázisokat is épít. A különböző forrásokból érkező heterogén adatok egy integrált rendszerbe migrálásával a *Huntéka-M* már nemcsak a múzeumi nyilvántartás funkcióit látja el, hanem szakirodalmi tudásbázisként képes kiszolgálni a kutatói igényeket is. A projekt szempontjából különösen jól használható a *Magyar írók bibliográfiája*[16] és a *Budapest topográfia*[17] az első számos Móricz-vonatkozású cikk leírását és forrásait tartalmazza, a másodikban pedig nyomon követhetjük Móricz Zsigmond budapesti lakcímeit.

[15] Ungváry Rudolf és Cserbák András, szerk., „Geotaurusz és Geohistaurusz: Földrajzi nevek és humángeográfiai nevek tezaurusza," 2001. nov. 1., hozzáférés: 2018.04.12, http://mek.oszk.hu/00000/00070/html/.

[16] *Magyar írók bibliográfiája*, hozzáférés: 2018.04.12, https://opac-adattar.pim.hu/.

[17] *Budapest topográfia*, hozzáférés: 2018.04.12, https://opac-nevter.pim.hu/.

 DOI 10.31400/dh-hun.2018.1.227




**3. ábra**. Egy Móricz-vonatkozású cikk rekordja a *Magyar írók bibliográfiája* adatbázisban. Forrás: `https://opac-adattar.pim.hu/record/-/record/PIM1367718`


**4. ábra**. Móricz Zsigmond lakcímének rekordja a *Budapest topográfia* adatbázisból Forrás: `https://opac-adattar.pim.hu/record/-/record/PIM1644715`

A *Huntéka-M* szabványos kimenettel és szabványos adatcsere-formátummal (HUNMARC) rendelkezik, ezért könnyen kommunikál más integrált rendszerekkel, így a DigiPhil bibliográfiai és besorolási adatait tároló *Koha*val is. A biblio.digiphil.hu mögött működő könyvtári komponens nyílt forráskódú, amint a DigiPhil más célszoftverei is. A *Koha*ban épülnek a Móricz-levelezéshez tartozó elsődleges és másodlagos bibliográfiák, és ide integrálódnak a más rendszerekből érkező besorolási állományok.

A Móricz-levelezés a bibliográfiai rekordok közül a kézirat, könyv, periodika, cikk, a besorolási rekordok közül pedig a személynév, földrajzi név, egységesített cím űrlapjait használja majd.

Az adatbázisban viszont problémát jelent egy elvi mű és a kiadások kapcsolatainak leképzése a MARC korlátozottsága miatt. A kritikai kiadás mellett épülő bibliográfiában egy adott mű rekordjában jelenne meg az is, ha egy regényből átdolgozás (pl. színdarab) készült, ahogy a Móricz által írt művek esetében ez többször előfordult. Szintén nehéz MARC-sémával leírni, amikor a levélben csak általánosan említenek egy művet, és nem egy konkrét kiadásról van szó, vagy amikor a mű címe csak

ötletként merül fel, de később nem íródott meg. MARC-ban az egységesített cím (a130) besorolási rekord almezőibe nem lehet elhelyezni a szerzőséget, azt csak a kapcsolódó, a mű kiadásait leíró bibliográfiai rekordok mutatják (c100 – Személynév főtétel). A megoldást a könyvtári világban egyre inkább teret hódító FRBR-alapú RDA (Resource Description and Access) katalogizálási szabályzat[18] jelentheti. A bibliográfiai tételek funkcionális követelményeit (FRBR) megfogalmazó entitáskapcsolat-modell külön értelmezi egy mű kifejezési formáját *(expression),* megjelenési formáját *(manifestation)* és példányát *(item).*[19]

A *Koha* wiki oldalán[20] láthatjuk, hogy a közösség fejlesztői már kidolgozták, hogyan igazítható a MARC-alapú rendszer az RDA igényeihez, és vannak már olyan projektek, amelyek sikeresen implementálták az FRBR rendszerét a *Koh*ába. Valószínűleg a DigiPhil előtt álló egyik fejlesztési feladat a projektet kiszolgáló adatbázis felkészítése lesz az RDA-alkalmazásra.

A *Koha* előnye, hogy rendelkezik Z39.50 protokollal, így képes más adatbázisokból rekordokat fogadni. A kritikai kiadáshoz a HUMANUS[21] Móricz-vonatkozású cikkei kerültek az adatbázisba, ahol az átvett rekordok 040-es mezője mutatja az eredeti forrást.

A Móricz-projekten belül a *Koh*ában tárolt bibliográfiai rekordoknak kettős funkciója lesz: egyrészt segítik a kutatási munkát, másrészt a DigiPhil oldalán tájékoztatnak a művek és a kéziratok metaadatairól. A kutatás feladatának tekinti a Móricz-bibliográfia építését – többek között ezért is értelmezik egységesített címként a szövegekben előforduló Móricz-műcímeket. Ha konkrét kiadásról vagy példányról esik szó egy levélben, akkor a szövegkódolásban ezt szintén jelzik.

### 3.2. Keresés

A metaadatokba és a szövegekbe illesztett egyedi azonosítók (vagyis az összetett struktúrájú TEI-elemkészlet alkalmazása) többfunkciós kereséseket tesznek majd lehetővé az infrastruktúráját kiaknázva a digitális kiadású DigiPhil Móricz-levelezésben.

A DigiPhil a kutatás során új keresőfelületet fejleszt, amely ötvözi a szabad szavas keresést és az XML-nyelv adta lehetőségeket. Egy indexelő alkalmazás a korpusz szövegének egészét feldolgozza, a szabad szavas keresésen túl lehetőség nyílik a csonkolt szavak és az ún. *joker* karakterek alkalmazására is. A keresőfelület másik oldalán az XML-ek hierarchiáját és elemkészletét kezelő eszköz áll. Ennek segítségével az egyes TEI-elemekre külön-külön is lehet keresni (például a törölt szövegrészekre: `<del>`); illetve különböző szűrési feltételeket lehet majd beállítani, így például ha dátumra keres a felhasználó, előre megadható lesz, hogy a háromféle datálást milyen sorrendben vegye figyelembe a keresőrendszer.

[18] „Resource Description and Access (RDA),” Library of Congress, hozzáférés: 2018.04.12, `http://www.loc.gov/aba/rda/`.

[19] „Functional Requirements for Bibliographic Records,” IFLA, hozzáférés: 2018.04.12, `https://archive.ifla.org/VII/s13/frbr/frbr_current_toc.htm`.

[20] „Koha: RDA,” hozzáférés: 2018.04.12, `https://wiki.koha-community.org/wiki/RDA`.

[21] Humántudományi Tanulmányok és Cikkek Adatbázisa, hozzáférés: 2018.04.12, `http://www.oszk.hu/humanus/`.

 DOI 10.31400/dh-hun.2018.1.227

## 4. A projekt összegzése

Az elmúlt másfél évben a kutatócsoport a projekt sikeres megvalósításához a szükséges és nélkülözhetetlen alapokat rakta le. Megtörtént a lelőhelyek feltérképezése, a kiadás szempontjából releváns, jelenleg elérhető forrásanyag számbavétele. Rendelkezésre állnak a PIM gyűjteményében található Móricz-levelezés példányairól készült fakszimilék, a képek szabályos elnevezése, ezen túlmenően a képszerkesztés folyamata befejeződött. A más közgyűjteményekben található levelek digitális másolatainak megrendelése folyamatban van.

Elkészült az 1913-ig keletkezett kéziratok főszövegeinek leírása (közel 1300 levél): a szövegkritikai elvek szerinti betűhív átiratok, valamint az emendálásokat tartalmazó olvasószöveg előállítása is. A szövegek ellenőrzése, összeolvasása, szoros időrendbe rendezése folyamatosan halad. A szerkesztőbizottság kialakította a levelekben előforduló szövegjellemzők és szövegműveletek jelölésére használt TEI-elemkészletet. A definiálni kívánt entitások azonosítása a személyneveket, földrajzi neveket és az egységesített címeket érinti, amelyek mindegyikéhez egyedi azonosítókat rendelnek a kereshetőség és a szemantikus kapcsolatok kiépítésének érdekében.

A munkatársak elvégezték a levelek formai feltárását, vagyis a kéziratok fizikai adatainak felvételét, valamint a szövegközlést a kritikai kiadásnak megfelelő részletességgel.[22] A kiadáshoz használt besorolási adatok tárolása és újabb rekordokkal való bővítése a *Koha*-rendszerben történik, ahol még a rekordstruktúra és a rekordkapcsolatok rendszere folyamatos fejlesztés alatt áll.

A projekt felénél, 2017-ben már számos tanulságot vonhatott le a kutatócsoport a munkafolyamatok eredményességét illetően. A szövegfeldolgozás egyes fázisainak munkamódszerei változó hatékonyságúnak bizonyultak. Bár a levelek szkennelése, a szövegek metaadatolása és *Microsoft Word*ben való jelölése megfelelően haladt, mára bebizonyosodott, hogy a köztes platform használata túl sok hibalehetőséget rejt a szövegtranszformáció során. A *Microsoft Word*-fájlokból nem lehet egy lépésben TEI XML-fájlokat kinyerni, csak bonyolult, többlépcsős folyamattal. A *Microsoft Word*ben történő átalakítás Visual Basic-kóddal (Visual Basic for Applications)[23] és reguláris kifejezések segítségével történik, majd a kinyert (még nem hierarchikus) XML-fájlokat *Oxygen XML Editor*ban[24] alakítják TEI XML-kóddá a projekt számára írt egyedi stíluslap segítségével. A *Microsoft Word*ből való átmásolás nehézségekkel terhelt a karakterkódolás miatt is (idézőjelek, rövid és hosszú kötőjelek keveredése), nem beszélve a szövegbevitel és a kódolás során történő hibás jelölésekről, gépelési hibákról. A többféle ellenőrzőprogram futtatása, a hibajavítások, az újabb ellenőrzések beiktatása mind jelentősen megnöveli a szövegtranszformációra fordított időt, és fennáll az adatvesztés veszélye. A projekt következő szakaszában a szövegek kódolása csak szabványos TEI XML-környezetben történhet a célnak megfelelő eszközzel. A

[22] Magyar Tudományos Akadémia I. osztályának Textológiai Munkabizottsága, „Alapelvek az irodalmi szövegek tudományos kiadásához," hozzáférés: 2018.04.12, `http://textologia.iti.mta.hu/alapelvek.pdf`.

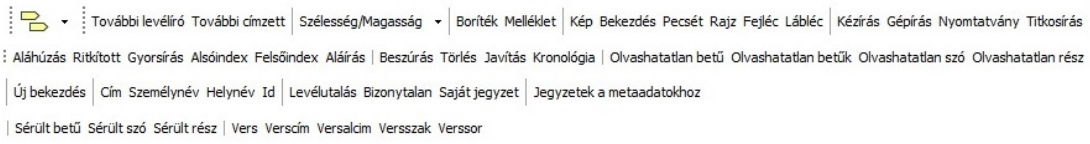
[23] „Visual Basic Guide," Microsoft, hozzáférés: 2018.04.12, `https://docs.microsoft.com/hu-hu/dotnet/visual-basic/`.

[24] *Oxygen XML Editor*, hozzáférés: 2018.04.12, `https://www.oxygenxml.com/`.

hibák kiküszöbölésére a DigiPhil csapata új leírókörnyezetet fejlesztett az *Oxygen XML Editor* programban, amely felváltotta a *Microsoft Word*öt mint adatbeviteli felületet.

A DigiPhil a virtuális kutatókörnyezetet[25] először az Arany János levelezése kritikai kiadásainak digitalizálási projektjén tesztelte, amely az Arany János Összes Művei 15–19. kötetekben található. A sikeres próbaidőszak után a Móricz-kutatócsoport is áttért az új környezet használatára a levelek leírásához. A kutatókörnyezet a *Microsoft Word*-del szemben számos előnnyel rendelkezik. Azon túl, hogy szabványos kimenetet biztosít, számos hibalehetőséget is megelőz. A kutatókörnyezet szintaktikai ellenőrző algoritmusokat tartalmaz, amelyek figyelmeztetnek a formalizálható szintaktikai hibákra. A DigiPhil a *Microsoft Word*-fájlok átalakításából átörökítette a köztes, kevés hierarchiát tartalmazó XML-struktúrát a levelek leírásához, mivel a TEI szerkezete rendkívül bonyolult, ez a struktúra jelentősen megnehezítette volna a kutatói környezet kialakítását, az XML-ek megjelenítését és a szintaktikai ellenőrzést. Ezekből a „sík" (kevés hierarchiát tartalmazó) XML-fájlokból a DigiPhil stíluslap segítségével állítja elő a publikálásra szánt TEI XML-fájlokat. A kutatókörnyezet használatához elég az *Oxygen*-szerkesztőt egyszer telepíteni és importálni a leíráshoz fejlesztett komponenseket. Természetesen a levelek leírása során új jelenségek bukkanhatnak fel (például ritkított betűkkel írt szó vagy sérült papír miatt olvashatatlan szavak), amelyek kódolására új jelölőket kell bevezetni, illetve ezeknek a környezetbe való beillesztését a szükséges módosításokkal elvégezni. A kutatói környezet egy másik előnyös tulajdonsága, hogy minimalizálja az adatvesztés lehetőségét, és biztosítja a kutatócsoport számára, hogy különböző munkaállomásokon dolgozzanak, a környezet ugyanis összeköttetésben áll egy változáskövető szerverrel, amelyre csak szintaktikailag helyes fájlokat menthetnek. A DigiPhil meghatározott időközönként archiválja a fájlokat egy repozitóriumban, ahonnan a változáskövető szerver esetleges leállása esetén is visszaállíthatók a fájlok.

A kutatókörnyezet kialakításakor a DigiPhil figyelembe vette azt az igényt, hogy a környezetet felhasználóbarát, irodai szoftvereket imitáló grafikus megjelenítéssel lássa el, amely hasonlít a már megszokott *Microsoft Word*-környezethez (menürendszer, magyar feliratú gombok), a gombok segítségével a megfelelő XML-jelölők automatikusan a kijelölt szöveghelyre kerülnek, így elkerülve a *Microsoft Word*-re jellemző szintaktikai hibákat.

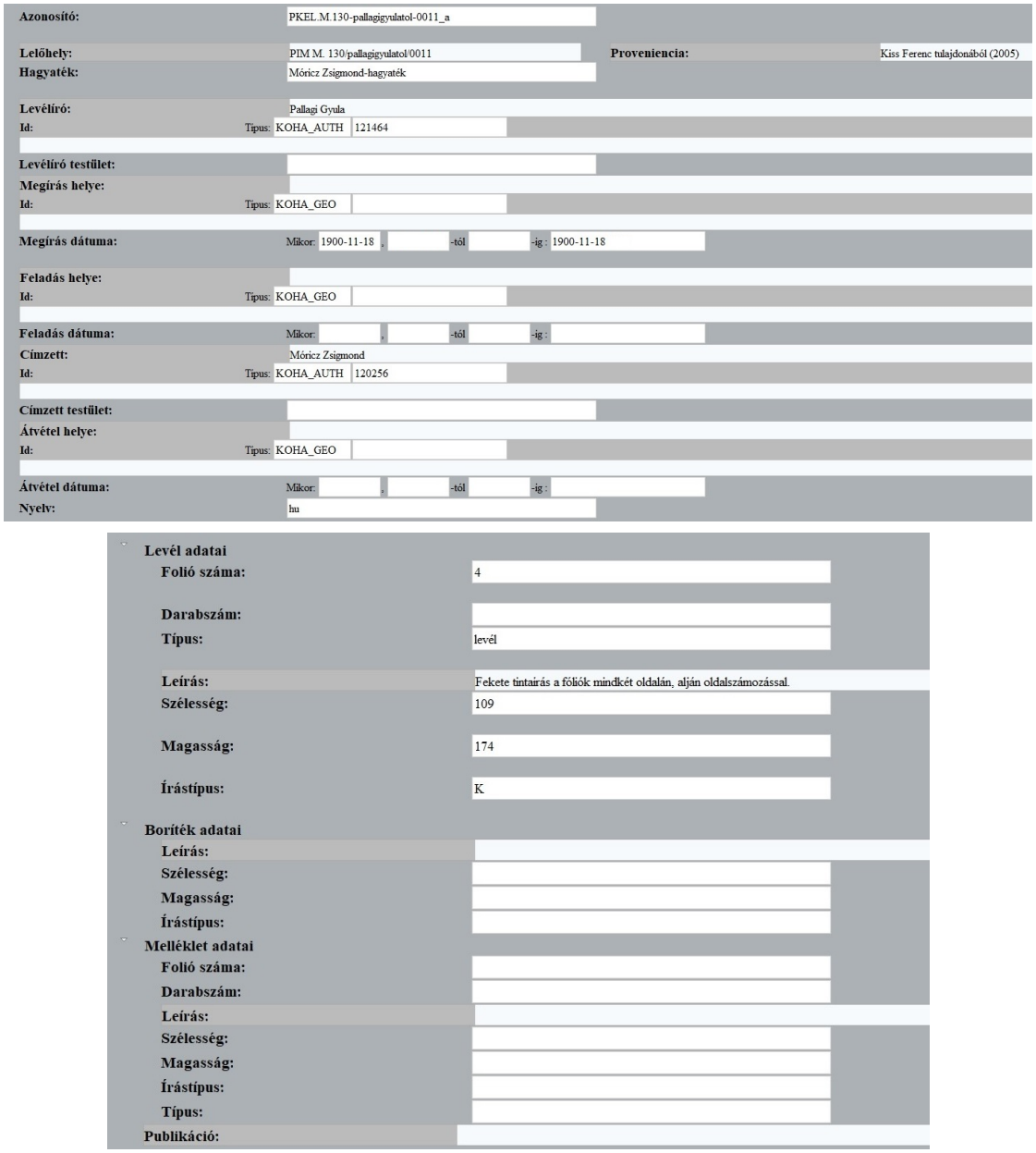



**5. ábra**. Magyar nyelvű menüsor a Móricz-kutatócsoport által használt leíró környezetben

A kutatócsoport a metaadatokat egy előre meghatározott mezőkkel rendelkező táblázatban adhatja meg:

[25] Palkó Gábor, „A digitális bölcsészet kultúrtechnikái. Virtuális kutatókörnyezetek," előadás *A humán tudományok és a gépi intelligencia* c. konferencián, Budapest, 2017. november 20.

 DOI 10.31400/dh-hun.2018.1.227



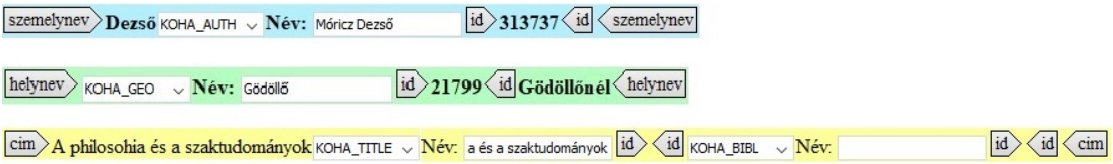

**6. ábra**. Móricz-kutatócsoport által kitöltendő táblázat

Bár a kutatócsoport már a *Microsoft Word*ben való leírás során is végzett adatgazdagítást, az automatikus átalakításhoz kidolgozott szintaxis rendkívül bonyolultnak bizonyult, a legtöbb szintaktikai hiba, elgépelés itt adódott, ami jelentős mennyiségű utólagos ellenőrzést és javítást igényelt a szövegtranszformáció elvégzése után.

Személy- és helynevek azonosítása, valamint bibliográfiai adatok kódolása *Microsoft Word*ben:

```
[személy] [@ Móricz Dezső @ KOHA_AUTH:313737] Dezső [személy vége]
[hely@ Gödöllő @] KOHA_GEO:21799 Gödöllőnél [hely vége]
[cím@ Hét krajcár@] KOHA_TITLE:3081083 KOHA_BIBL:40125836 Hét
krajcár[cím vége]
```

A megfelelő elem beillesztése után a kutatókörnyezet automatikusan létrehozza a kitöltendő mezőket, illetve előre kitölti a névterek DigiPhilben használt azonosítóját. A rendszer által javasolt azonosító (KOHA_AUTH) a személynevek esetén legördülő ablakban jelenik meg, a kutatócsoport itt választhat más névteret (PIM, LOK).




7. **ábra**. Személy- és helynevek azonosítása, bibliográfiai adatok kódolása kutatói környezetben

Az új kutatókörnyezet kialakításán túl módosult a levelek (és a hozzájuk tartozó XML-fájlok) ellenőrzésének folyamata is. Míg a korábbi tervek szerint a *Microsoft Word*ben kódolt levelek transzformációja után a javítás a TEI XML-fájlokban zajlott volna, a jövőben a DigiPhil a kutatócsoport számára egy olyan tesztoldalt biztosít, amely a leíró környezetben leírt „sík" XML-eket jeleníti meg, az ott megszokott vizualizációval. Ennek a felületnek a segítségével a már leírt levelek eljuttathatók azoknak a szakértőknek is, akik nem vettek részt a levelek átírásában: ők online, a kutatói környezet telepítése nélkül kapcsolódhatnak be a kutatásba. A tesztfelületen történő ellenőrzés után a kutatócsoport a javításokat még a kutatói környezetben végzi el, majd ezt követően alakítja át a DigiPhil a leveleket, és publikálja a hivatalos oldalán.

Arról, hogy milyen lehetőséget nyújt a digitális médium a kritikai kiadás számára, milyen vizsgálati módszereket ajánlhat fel a levélszövegek vizsgálatára, csak a 2019-ben lezáruló kutatási időszak után nyújtható részletesebb, elméleti kérdéseket is érintő összefoglaló.

## The Digital Critical Edition of the Correspondence of Zsigmond Móricz (1892–1913): a Case Study

The NKFIH-project that seeks to publish the digital edition of Zsigmond Móricz's (1892–1913) correspondence was launched in 2016 at the Petőfi Literary Museum in Budapest. The project itself has been a huge challenge for the Móricz-research group because they can only partially rely on the earlier paper-based edition. Drawing on the experiences and resolved problems of the first years of the project, this paper focuses on the harmonization and relationship of a museum's programming/IT possibilities, the philological problems and the applied software capacities. As the paper discusses the challenges of an ongoing project, the study does not offer a holistic and comprehensive overview of the entire project, rather a list of problems as encountered along the way and their solutions.

Keywords:
digital philology, Zsigmond Móricz, digital scholarly edition, correspondence, DigiPhil

 DOI 10.31400/dh-hun.2018.1.227