



AZ OKTATÁS, A KUTATÁS ÉS A KÖZGYŰJTEMÉNYEK DIGITÁLIS TRANSZFORMÁCIÓJA FELSŐFOKON

NETWORKSHOP 2024
33. Országos Informatikai Konferencia

2024. április 3–5.
Eszterházy Károly Katolikus Egyetem, Eger

AZ OKTATÁS, A KUTATÁS ÉS A KÖZGYŰJTEMÉNYEK DIGITÁLIS TRANSZFORMÁCIÓJA FELSŐFOKON

NETWORKSHOP 2024
33. Országos Informatikai Konferencia

2024. április 3–5.
Eszterházy Károly Katolikus Egyetem, Eger

Szerkesztette: Tick József, Kokas Károly, Holl András

HUNGARNET Egyesület
Budapest, 2024



HUN-REN
Magyar Kutatási Hálózat

NETWORKSHOP

Szerkesztette: Tick József, Kokas Károly, Holl András

Tipográfia és tördelés: Vas Viktória

Korrektúra: Danyi Melinda

Angol nyelvi lektor: Cseresnyés Dorottya

Networkshop 2024 konferencia előadásainak közleményei

Eszterházy Károly Egyetem, Eger

2024. április 3–5.

ISBN 978-615-82243-2-1

DOI: <https://doi.org/10.31915/NWS.2024>

Kiadja a HUNGARNET Egyesület
az MTA Könyvtár és Információs Központ közreműködésével

Budapest

2024

Borítókép: [freepik.com](https://www.freepik.com)

TARTALOMJEGYZÉK

Előszó	5
Ungváry Rudolf A MARC21 formátum kettős szerkezete és a formátum felhasználói szintjének fordításai	7
Holl András, Andódy Katalin Adatbányászati gyakorlatok repozitóriumra és MTMT-re.....	16
Simon András Mesterséges intelligenciával támogatott adatgazdagítás a Nemzeti Levéltárban.....	22
Soós Gábor, Rövid András, Ormos Pál V2X – A járművek közötti kommunikáció kihívásai	29
Csernai Zoltán Egy online tanulást támogató portál kurzusának vizsgálata Big Data adatelemző módszerekkel.....	36
T. Nagy László, Németh Áron A mesterséges intelligencia (MI) teológiai kompetenciái	45
Mészáros Erika Kodaktól a jövőig – Egy könyvtári digitalizálás szinterei.....	55
Frankó Máté, Sándor Ákos Adatvizualizáció a könyvtári menedzsmentben: fejlesztések az SZTE Klebelsberg Könyvtár döntéstámogató rendszerében.....	63
Hernek István A felhasználóképzés szintjei az SZTE Klebelsberg Könyvtárban: az elsőévesektől a kutatókig	72
Némethi-Takács Margit, Borbély Mária Bibliográfiai kapcsolatok az általános megjegyzés adatmezőben.....	78
Dobás Kata, Tüskés Anna A magyar irodalomtörténet bibliográfiájának migrációja az ITIdata szemantikus adatbázisba	87
Horváth Péter A kanonikus magyar költészet versformakeresője.....	96
Sebestyén Ádám, Sárközi-Lindner Zsófia Történeti források szemantikus feldolgozása – Az ELTEdata adatbázis új gyűjteményei	105

Bolya Mátyás

Lyukkártya és népdalrendezés – Egy mechanikus népzenei adatbázis digitális rekonstrukciójának lehetőségei.....112

Kovácsházy Tamás

Az idő, mint alapvető infrastruktúra, az idő szerepe az adatközpontban.....121

Albert Ágota Katalin

A mesterséges intelligencia használatának követelményei az oktatási szektorban, különös tekintettel a mesterséges intelligencia használatáról szóló rendelethez.....129

Varga Emese

Digitális szövegszerkesztés a dHUpa keretrendszerében135

Nemoda Zsuzsanna, Héjja Balázs, Nagy Andor, Tóth Máté

A Pest Megyei Digitális Könyvtár fejlesztése 141

Nagy Dóra, Sándor Ákos

Voice2text: a hanganyagátírás lehetőségei MI segítségével.....149

Kalcsó Gyula

Képek és metaadataik gyűjteményezése scrapingtechnológiával közösségi képmegosztó oldalról157

Péter Róbert, Szántó Zsolt, Biacsi Zoltán, Kocsis Zoltán, Berend Gábor, Bilicki Vilmos

Az AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) többnyelvű kutatási eszköz munkafolyamata és új funkciói163

Máray Tamás

Kvantum-számítástechnika: ez már a „jövő”?171

Fellegi Zsófia

Digitális kiadások migrációja: gépi és emberi intelligencia együttműködése.....177

Palkó Gábor

Posztmodern intertextualitás és digitális szövegkiadás184

Antal Dániel

A szlovák adatkicserélési tér magyarországi föderációjának lehetőségei.....192

Vass Johanna

Kutatási adatok megosztása a gyakorlatban – Adatrepozitóriumok használata az Ökológiai Kutatóközpont publikációiban199

Mihály Eszter, Micsik András, Nagy Kadosa

Irodalmi levélváltások nyomában TEI-vel és térképpel.....208

Történeti források szemantikus feldolgozása – Az ELTEdata adatbázis új gyűjteményei

Semantic data processing of historical sources – New collections of ELTEdata

Sebestyén Ádám
Digitális Örökség Nemzeti Laboratórium
sebestyen.adam@btk.elte.hu

Sárközi-Lindner Zsófia
Digitális Örökség Nemzeti Laboratórium
slzsofia@student.elte.hu

Absztrakt

Az ELTE BTK Digitális Örökség Nemzeti Laboratóriumban fejlesztjük wikibase-alapokon működő szemantikus adatbázisunkat, az ELTEdatát. Folyamatosan dolgozunk mind az újabb kutatási anyagok és gyűjtemények integrálásán, mind a meglévő funkciók fejlesztésén, adott esetben új kiegészítések telepítésén. Jelen tanulmány az újonnan elért fejlesztések és eredmények bemutatására vállalkozik, a legutóbbi, nagy mennyiségű adatbetöltésre helyezve a hangsúlyt, melynek keretében a kora újkori Győri Jezsuita Gimnázium anyakönyvéből kiadott adattár névanyagát dolgoztuk fel. Az adatfelvitel egy tesztanyag manuális létrehozását követően félautomatikus módon történt, eleinte a QuickStatements Wikidata-eszköz, illetve API-n keresztül, egy erre a célra írt Python-kód segítségével. A tanulmány az elért eredmények mellett az automatizálás során felmerült nehézségeket is számba veszi. A bizonytalan adatok nagy száma, a változó értékek (például a települések és személyek névváltozatai) leképezése és összehangolása jelentették az automatikus betöltés előtti legnagyobb kihívásait.

Kulcsszavak: szemantikus adatbázis, Wikibase, félautomatikus adatgazdagítás

Abstract

ELTEdata, developed by the National Digital Heritage Laboratory of Eötvös Loránd University, aims to organize the sources of prosopographical, bibliographical and other historical research groups into a semantic data network. This paper presents our new developments, focusing on the recent entry of a large amount of data. Our primary source was the register of the Jesuit Gymnasium of Győr, containing data of matriculation from the 17th and 18th centuries. Data entry was accomplished after the specification and the creation of a test material semiautomatic, using the tool QuickStatements (decisively in the first phase of the workflow) and API, by means of a code, written in Python. The paper takes into account our current results and the difficulties during the automation as well. The uncertain data, the diverse forms of values (aliases, for instance) proved to be the greatest challenge during the creation of the namespace.

Keywords: semantic database, Wikibase, semiautomatic data enrichment

Az ELTEdata szemantikus adatbázis egy wikibase-szoftveren alapuló szolgáltatás, amit az ELTE Digitális Bölcsészet Tanszék, valamint a Digitális Örökség Nemzeti Laboratórium fejleszt történeti tárgyú kutatások anyagainak szemantikus adathálózatba rendezésére és közzétételére.¹ Az utóbbi évek Networkshop-konferenciáin (2021-ben és 2022-ben) e tanulmány első szerzője részletesen bemutatta az adatbázis funkcióit és addigi gyűjteményeit, külön kitérve olyan eszközök ismertetésére, mint a SPARQL szemantikus lekérdezőfelület és a félautomatikus adatbevitelt lehetővé tevő QuickStatements.² Az ELTEdata fejlesztése jelenleg is tart: egyrészt telepítve lettek kiegészítő modulok, másrészt új kutatási anyagok entitásaival bővült a névtér. Említést érdemelnek még az ELTE modern magyarországi történelemmel foglalkozó doktoranduszainak formálódó gyűjteményei is. A doktori program hallgatói a 2023-as őszi félévben egy szeminárium keretei közt ismerkedhettek meg az ELTEdata adatbázissal, betekintést nyerve az adatfelvitelbe és ontológiaépítésbe, valamint a SPARQL használatába. PhD-workshopokon több alkalommal lehetőség nyílt bemutatni a szolgáltatást, így módon fontos távlati cél, hogy az ELTEdata, illetve a szemantikus adatbázisok alapszintű ismerete és használata a jövőben is része legyen a digitális bölcsészettel kapcsolatos képzéseknek.

Az új projektek közül kiemelkedik a győri jezsuita gimnázium adattára, mintegy 20000, javarészt félautomatikus módon betöltött entitással.³ A projekt munkatársai körülbelül egy éve kezdtek el intenzíven foglalkozni a forrásanyaggal, konzultálva az adattár papíralapú, illetve digitális feldolgozását végző kutatócsoporttal. A kutatócsoport még 2015 és 2019 között, a *Katolikus iskoláztatás a kora újkori Magyarországon* NKFI-projekt keretében, Fazekas István vezetésével kezdte meg a győri gimnáziumi anyakönyv feldolgozását és kiadását.⁴ Ennek folytatását jelentette az a 2019-ben kezdődött kutatás, mely a forrásanyag elemzését, a diákság társadalmi hátterének feltárását tűzte ki célul.

Az adattár az 1630 és 1773 között a győri jezsuita gimnáziumba beiratkozott diákok életrajzi és tanulmányokkal kapcsolatos adatait rögzíti. A közvetlen forrásunk egy Excel táblázat volt, amiben minden sor egy beiratkozást jelölt. Ez az oka annak, hogy a táblázatban ugyanahhoz a személyhez több sor is tartozhat. A leképezés során a diákok nevei lettek az entitások, így ezek összehangolása a különböző beiratkozásokkal több problémát is felvetett, melyekről részletesen a későbbiekben lesz szó. Az automatizálás egyik fő nehézségének a számtalan alakváltozat bizonyult. A családi nevek például sok esetben beiratkozásokként változnak, a kutatócsoport ugyanis a forrásban szereplő névváltozatokat vette alapul, csak a keresztnéveket egységesítették. A személyek megkülönböztetését (rengeteg az azonos család- és keresztnévvel rendelkező diák) azok egyedi azonosítója teszi lehetővé. A személyekre vonatkozó életrajzi adatok közül az első a származási hely: a táblázat H oszlopa (lásd az 1. ábra) a katalógusban használt latin alakot közli, az I oszlop pedig az ebből kikövetkeztetett mai magyar településnevet. Ide tartozik még, mintegy tágabb föld-

1 Az ELTEdata nyitóoldala az eddig véglegesített, kész projektekkel:
https://eltdata.elte-dh.hu/wiki/Main_Page (Hozzáférés dátuma: 2024. 06. 19.)

2 Sebestyén Ádám, „Az ELTEdata szemantikus adatbázis legújabb fejlesztései”, in *Networkshop 2022*, szerk. Tick József, Kokas Károly, Holl András, 179–185 (Budapest: HUNGARNET Egyesület, 2022)

3 A névtér entitásait lásd az alábbi SPARQL-lekérdezésből generált linken: <https://tinyurl.com/2akg2av6>

4 *A győri jezsuita gimnázium diáksága. Anyakönyvi adattár (1630–1773)* szerk. Fazekas István, Kádár Zsófia, Kökényesi Zsolt, (Jézus Társasága Magyarországi Rendtartományának Levéltára: Budapest, 2020)

rajzi egységet jelölendő, a vármegye is. A *natio* a személyek etnikumára vonatkozik, és a forrásban használatos megnevezést követi, tehát egységesen latin alakban szerepel. Bár az esetek nagy részében egyértelműek, előfordul, hogy némelyik jelentése nem világos vagy kétértelmű: a Rascianus (Q16011, Q19788)⁵ például egyszerre jelöli a szerbeket és a görögkeleti egyház követőit. Ezekben az esetekben adott *natio*-entitás ELTEdata-oldalán, a label-ben magyaráztuk meg a szó jelentését. A társadalmi státusz és a vallás könnyen és egyértelműen leképezhető adatok, annyit viszont érdemes megjegyezni, hogy utóbbi nem mindig követi az excel-táblázat sztenderdizált, latin névalakjait: a nagy felekezetek, mint például a katolikus vagy a református, már létező entitások voltak az ELTEdata-ban, így ezeket felesleges és problémás is lett volna újból létrehozni. A specifikus, csak az adattár gyűjteményében előforduló felekezeteket természetesen a forrásban használt latin alakban képeztük le. A tanulmányokra vonatkozó adatokat két állítás írja le: az egyik a Győri Jezsuita Gimnázium, mint alma mater (ez természetesen az összes személynél ugyanaz), a másik az évfolyam. A beiratkozás éve minősítőként szerepel, adott évfolyamhoz rendelve (lásd a 2. ábra). A táblázatban megtalálható még a primer forrás (minden esetben a gimnáziumi adattár, *Catalogus Discipulorum Jauriensis* néven), a vonatkozó kötet és oldal-számmal kiegészülve, illetve, mint az a Q, az R és S oszlopokban látható, az esetleges megjegyzések, illetve a szakirodalom.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	SORSZ	AZONOS	AZONOS	CSNÉV	UNÉV	ELNÉV	SZÁRMHELY	MAIMAGYAR	MEGYE	NATIO	TÁRSÁLL	VALLÁS	ÉV	OSZTÁ	FORRÁS	MEGJEJYZE	MEGJEJYZE 2	IROK	
2	10000	10000	10000	Farkas	Andreas		Uyvariensis	Érsekújvár	Nyitra	Ung.				1630	POET	Cat. laur. I. 3r			Fuit viccomes comitatus Jaurinensis.
3	10247	10000	10000	Farkas	Andreas		Uyvariensis	Érsekújvár	Nyitra	Ung.				1631	RHET	Cat. laur. I. 3v			
4	10562	10000	10000	Farkas	Andreas		Uyvariensis	Érsekújvár	Nyitra	Ung.				1632	RHET	Cat. laur. I. 34r			
5	10001	10001	10001	Suponitz	Georgius		Jaurinensis	Győr	Győr	Ung.				1630	POET	Cat. laur. I. 3r			
6	10252	10001	10001	Suponitz	Georgius		Jaurinensis	Győr	Győr	Ung.				1631	RHET	Cat. laur. I. 3v			
7	10003	10003	10003	Szalay	Joannes		Somogyensis	Somogy m.	Somogy	Ung.				1630	POET	Cat. laur. I. 3r			
8	10201	10003	10003	Szalay	Joannes		Somogyensis	Somogy m.	Somogy	Ung.				1631	RHET	Cat. laur. I. 3v			
9	10004	10004	10004	Jantolovich	Joannes		Varadensis	Varasd	Varasd	Cro.				1630	POET	Cat. laur. I. 3r			
10	10259	10004	10004	Jantolovich	Joannes		Toplicensis	Toplica	Varasd	Cro.				1631	RHET	Cat. laur. I. 3v			
11	10006	10006	10006	Gaspartz	Joannes					Cro.				1630	POET	Cat. laur. I. 3r			
12	10294	10006	10006	Gaspartz	Joannes		Crapinensis	Krapina	Varasd	Cro.				1631	RHET	Cat. laur. I. 3v			
13	10007	10007	10007	Suchla	Joannes		Egbellensis	Egbell	Nyitra	Slav.				1630	POET	Cat. laur. I. 3r			
14	10258	10007	10007	Suchla	Joannes		Egbellensis	Egbell	Nyitra	Slav.				1631	RHET	Cat. laur. I. 3v			
15	10008	10008	10008	Szentmihály	Joannes					Transyl.				1630	POET	Cat. laur. I. 3r			
16	10257	10008	10008	Szentmihály	Joannes		Transylvanus	Erdély	Erdély	Ung.				1631	RHET	Cat. laur. I. 3v			
17	10009	10009	10009	Krisinus	Joannes					Cro.				1630	POET	Cat. laur. I. 3r			
18	10253	10009	10009	Crismus	Joannes					Cro.				1631	RHET	Cat. laur. I. 3v			
19	10010	10010	10010	Torkos	Joannes		Jaurinensis	Győr	Győr	Ung.				1630	POET	Cat. laur. I. 3r			
20	10290	10010	10010	Torkos	Joannes		Jaurinensis	Győr	Győr	Ung.				1631	RHET	Cat. laur. I. 3v			
21	10562	10010	10010	Torkos	Joannes		Jaurinensis	Győr	Győr	Ung.				1632	RHET	Cat. laur. I. 34r			

1. ábra: Az adatbetöltéshez használt közvetlen forrás

Ahogy az már szóba került, a forrásként használt táblázatban a beiratkozás, és nem a személy az alapegység, tehát egy sorban egy beiratkozás adatai jelennek meg. Mivel az egyes években különbözhet a bejegyzett családnevek alakja, ezért a sorokat az egyedi azonosító kapcsolja össze. Ez azért is hasznos, mert a gyakori névazonosságok miatt egyszerűbb az azonosítón keresztül eljutni adott személyhez, amire egy SPARQL-lekérdezés tökéletesen alkalmas. A sokféle alakváltozat miatt azt az elvet követtük, hogy az entitás label-jén mindig a név első előfordulása van feltüntetve, míg a többi alakváltozat sztring formában szerepel az állítások között, sorrendjük pedig a beiratkozások sorrendjét követi.

A tanulmányokra vonatkozó adatok leképezése előtt érdemes néhány szóban ismertetni a korabeli jezsuita oktatási modellt. 1599 óta a jezsuiták egyetemes oktatási rendjét a *Ratio Studiorum* című kézikönyv határozta meg. Nem volt ez másképp a győri gimnázium esetében sem. Ez az oktatási modell öt gimnáziumi évfolyamot különböztetett meg: három grammatikai osztályt (*principia, grammatica, syntaxis*), valamint a *humaniora* két felsőbb

5 A továbbiakban a példaként idézett entitásoknál és tulajdonságoknál feltüntettem azok ELTEdata-azonosítóit.

évfolyamát (*poetica, rhetorica*).⁶ Elvben a gimnáziumi felvétel követelménye volt, hogy a diákok rendelkezzenek alapvető ismeretekkel (anyanyelvi írás-olvasás, alapfokú számtan), a gyakorlat azonban azt mutatja, hogy ezek hiányában is be lehetett iratkozni. Hosszú távon tehát szükség volt olyan alsóbb osztályokkal bővíteni a gimnáziumot, ahol a tanulók hozzáférhettek ezen alapismeretekhez. Ezt a célt szolgálta az úgynevezett *parvista* osztályok megalapítása. Az évfolyamok rendszerét is a Wikidata-adatmodellt követő tulajdonság–érték párok képezik le. A beiratkozás időpontja, mint kezdő dátum, minősítőként van hozzáadva az állításhoz. A hivatkozás alatt található a forrás (a gimnáziumi katalógus vonatkozó kötet- és oldalszámának megjelölésével), valamint az esetleges egykorú, illetve a kutatócsoport által hozzáfűzött megjegyzések (lásd a 2. ábra bal oldali képmellékletét).

A bizonytalan, illetve nehezen leképezhető adatok közül a származási helyeket kell először megemlíteni. Ahogy azt már korábban jeleztem, a származási helyeknél az alapadat mindig a katalógusba bejegyzett, egységesített alakban közölt latin név. Ebből az alakból következtették ki a kutatócsoport munkatársai a település mai magyar elnevezését, amennyiben ez lehetséges volt. Számtalan olyan eset akad ugyanis, amikor csupán a latin alak van feltüntetve, lévén, hogy annak azonosítása sikertelen. A leképezés során az entitás mindig a település mai névalakja: például Zalaegerszeg, míg a forrásban szereplő Egerszeg-alak névváltozatként van hozzáadva a település adatlapján. Ugyanott került feltüntetésre, melyik vármegyében található adott város vagy falu, a történelmi közigazgatási egység tulajdonság (P132) segítségével. Ez az adat tehát a személyek ELTEdata-oldalán nem látható, vagyis nem egy minősítővel lett a származási helyhez csatolva, hanem annak adatlapján jelenik meg. SPARQL-lekérdezővel így is lehet szűrni ezeket az adatokat, vagyis megoldható a személyek vármegyék alapján történő listázása. Ami a származási helyhez minősítőként csatolva jelenik meg, az a forrásban használt megnevezés (P163), vagyis a katalógusban feltüntetett latin névváltozat. Mivel a származási hely-entitás neve mindig a mai településnév, törekedni kellett arra, hogy az egységesítés ellenére megőrizzük az egyes névváltozatokat is. Példaként hozható fel az egykori pozsony-vármegyei Szentistvánkút (Q18607), ami a község mai magyar megnevezése (szlovákul Studienka). Van ugyanakkor egy korábbi szlovák neve is, mégpedig Hasprunka (magyaros formában Haszprunka). A forrásban ez a történelmi megnevezés szerepel az első helyen, lévén, ez áll legközelebb a latin alakváltozathoz. Fel van tüntetve ugyanakkor a szintén egykorú német neve is (Hausbrunn), ami tulajdonképpen az eredeti név, mivel ausztriai telepések alapították a 15. században (lásd a 2. ábra jobb oldali képmellékletén).

Nem ritka, hogy a forrásban szereplő latin alak alapján több lehetséges település is szóba jöhet. A „Tejediensis” alak megfejtése egyértelműen Tejed, jóllehet, maga a név három lehetséges települést is takarhat: Ollétejed és Előtejed ma Dunaszerdahely városrészei, egykor önálló falvak, Lidértejed pedig a mai Pódatejed településrésze, egykor szintén önálló falu. A leképezést ebben az esetben úgy oldottuk meg, hogy a korabeli vármegyét tüntettük fel egyértelmű származási helyként (mindhárom Pozsony vármegyében található), és egy lehetséges értékminősítővel (P162) vezettük fel a szóba jövő három községet. Egy másik megoldás, hogy a forrás alapján kikövetkeztetett mai nevet hozzuk létre entitásként, és ennek az oldalán tüntetjük fel a lehetséges értékeket, ahogyan ez például Csátár (Q21250) esetében megvalósult. A bizonytalan származási helyek szintén gyakori esete,

6 A győri jezsuita gimnázium diáksága, 14.

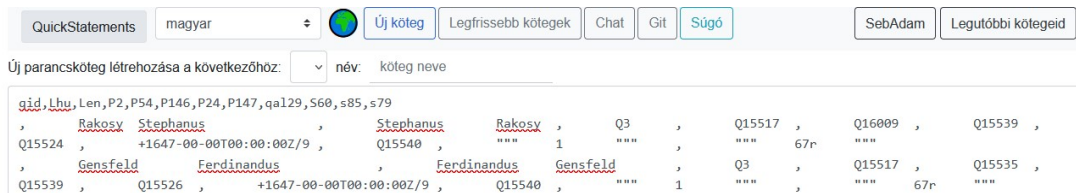
mikor a latin forrásból nem deríthető ki egyértelműen azonosítható település. Bizonytalan adat persze itt is megadható, legtöbbször azonban megközelítőleg sem lehet azonosítani a származás pontos helyét.

poetae		további feltüntetett név	Haszprunka
kezdő dátum	1630		→ 0 hivatkozás
▼ 1 hivatkozás			Haszprunka
állítási forrása	Catalogus discipulorum Jauriensis Gymnasii Societatis Jesu		→ 0 hivatkozás
kötet	1		Hausbrunn
oldalak	5r		→ 0 hivatkozás
megjegyzések	Fuit vicecomes comitatus Jaurinensis.		

2. ábra: Részletek a leképezett entitások közül

Az adatbetöltés kezdeti szakasza és a mintaanyag létrehozása a specifikáció kialakításával párhuzamosan, manuális módon történt. Hamar nyilvánvalóvá vált, hogy a félautomatikus betöltési lehetőségeket nagyban meghatározza a különféle alakváltozatok egységesítése. Mindenek előtt a származási helyeknél merült fel ez a probléma. Mivel egyedül a forrásban használt megnevezések egységesek, szükséges volt a lehetőségekhez képest minél jobban uniformizálni a települések mai névalakjait. Hogy a betöltéshez használt Python-kód számára egyértelmű legyen két különböző alakban használt helységnév lényegi azonos-sága, az összes település adatlapján fel kellett tüntetni az esetleges változatokat. Például, ami a táblázatban Pruzsina alakban szerepel, az ELTEadatában (magyar nyelvű nézetben) a mai hivatalos magyar névén, Barossháza alakban található meg (Q18505). A másik alakváltozat hozzáadásával azonban megoldhatóvá vált, hogy a táblázat félautomatikus feldolgozása során a kód megtalálja a cellában szereplő névhez tartozó label-t. Hasonló kérdések merültek fel a változó értékekkel kapcsolatban. Gyakran előfordul, hogy különböző években különböző náción, vagy épp felekezet van rögzítve a diákok beiratkozásainál, de az sem ritka, hogy a társadalmi státuszban következik be változás. Az ilyen helyzetekben jónak láttuk egy minősítővel évszámot csatolni az állításhoz, jelezve, hogy adott személyre ez az állítás ebben az évben volt érvényes.

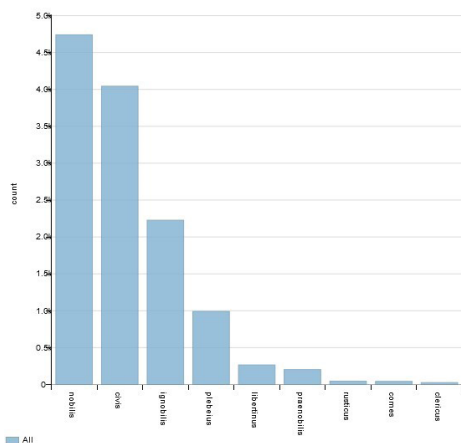
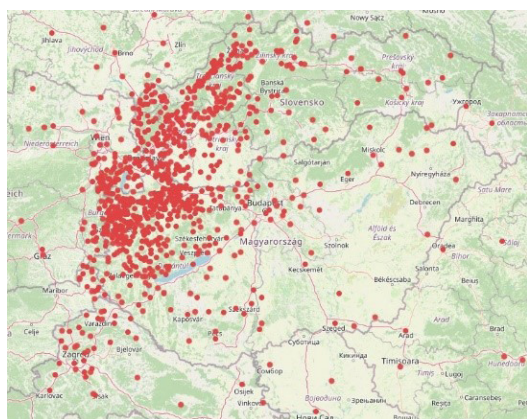
A félautomatikus adatbevitelt a QuickStatements-eszközzel kezdtük el, túlnyomórészt az egyszerű és egyértelmű adatokat dolgozva fel a segítségével. A táblázatban több ezres nagyságrendben szerepeltek nevek, ahol csak egy beiratkozásáról tudunk (tehát egy név egy sor), ráadásul csupán a beiratkozásuk időpontja és az osztályuk ismert, a náción, a felekezeti és a társadalmi hovatartozás nem. Ezek közül megközelítőleg 5000 személyt került be QuickStatements-szel az adatbázisba. Az Excelből viszonylag könnyen lehetett dolgozni, de a header kialakításánál figyelembe kellett venni, hogy a tulajdonság-érték párok mellett minősítők és hivatkozások is szerepelnek, tehát ennek megfelelően kell kiválasztani a szükséges prefixumokat. Ide tartozik még, hogy a formátum sajátosságából adódóan figyelni kellett például arra, hogy a sztring adattípusú értékek három idézőjelben legyenek feltüntetve.



3. ábra: Az ELTEdata QuickStatements-betöltőfelülete

A táblázat legnagyobb része egy Pythonban írt kóddal, API-n keresztül lett betöltve az ELTEdata felületére. Ezt a munkát Sárközi-Lindner Zsófia végezte egy Python-könyvtár, a Wikidataintegrator segítségével.⁷ Ezen a módon alapvetően a Wikidatába lehet nagy mennyiségű, strukturált adatot betölteni, de más wikibase-rendszerekhez is használható. Az adatbetöltést megelőző munkák nagyrészt a Pythonban végzett műveletek voltak, egy egységes mintázat kialakítása érdekében, amit a kód használni tud. Így például a táblázat beolvasása, rendszerezése, a „többsoros”, tehát többször beiratkozott személyek adatainak összeszedése, azok formázása és a zajos adatok kiszűrése tették ki a munkafolyamat nagy részét. A kétféle félautomatikus betöltési móddal a táblázat majdnem összes személye bekerült az adatbázisba. Kisebb mértékben maradtak olyan nevek, amiket a kód nem tudott olvasni, jelenleg ezek pótlása zajlik. Hátra van még továbbá a minősítők ellenőrzése, illetve egységesítése. A többszöri beiratkozás miatti változásoknál több helyen maradtak olyan állítások, ahol a dátum-minősítő lemaradt, ezeket mindenképp érdemes volna pótolni.

Noha a manuális utómunkák, javítások még szükségesek, az eddig betöltött adatmennyiség reprezentatívnek minősül, ahogy ezt a SPARQL-lel generált adatvizualizációk is igazolják. A származási helyek földrajzi vizualizálásával jól kirajzolódik az a sáv a történelmi Magyar Királyság nyugati részén (mindenek előtt Győr és Nagyszombat vonzáskörzettel), ahonnan a diákok felkeresték a jezsuita gimnáziumot a 17–18. században.



4. ábra: Adatvizualizációk a betöltött anyagon

⁷ A könyvtárhoz tartozó részletes leírás az alábbi linken érhető el: <https://pypi.org/project/wikidataintegrator/>

A gyakoriságot vizsgáló aggregátor funkció segítségével lekérdezéseket lehet írni többek közt a nációk, felekezetek vagy épp a társadalmi rétegek megoszlására. Legalább ennyire fontosak az olyan tabuláris jellegű adatvizualizációk, melyek a szemantikus formában feldolgozott adatokat szűrik, illetve rendszerezik. Annak megválaszolása, hogy adott évben hányan és kik iratkoztak be a gimnáziumba, SPARQL-lel meglehetősen egyszerű. Egy ilyen alapkérdezést aztán tetszőlegesen lehet bővíteni az osztályokkal, mi több, opcionális adatként a személyek etnikai, felekezeti és társadalmi hovatartozása is szűrhető. Ily módon olyan lekérdezések és vizualizációk generálhatóak, melyek alapján összevethető, hogy adott években milyen arányban változott a beiratkozottak felekezeti vagy etnikai hovatartozása. Összességében tehát a kutatócsoport eredeti kiadványában közzétett statisztikák is bővíthetőek, kiegészítve általuk kevésbé vizsgált szempontokkal. Mindez nagyszerűen megvilágítja (a nehézségek ellenére is) a szemantikus adatfeldolgozás előnyeit olyan történelmi források esetében, mint például egy kora újkori gimnáziumi anyakönyvi adattár.

Irodalomjegyzék

- Sebestyén Ádám, „Az ELTEdata szemantikus adatbázis legújabb fejlesztései”, in *Networkshop 2022*, szerk. Tick József, Kokas Károly, Holl András, 179–185 (Budapest: HUNGAR-NET Egyesület, 2022)
- A győri jezsuita gimnázium diáksága. Anyakönyvi adattár (1630–1773)* szerk. Fazekas István, Kádár Zsófia, Kökényesi Zsolt, (Jézus Társasága Magyarországi Rendtartományának Levéltára: Budapest, 2020)