

AZ ITIDATA SZEMANTIKUS ADATBÁZIS ÚJ KERESŐFELÜLETEI¹

Dobás Kata

HUN-REN Bölcsészettudományi Kutatóközpont Irodalomtudományi Intézet

dobas.kata@abtk.hu

ORCID: [0009-0009-7632-8276](https://orcid.org/0009-0009-7632-8276)

Szekrényes István

Eötvös Lóránd Tudományegyetem, Digitális Bölcsészeti Tanszék

Debreceni Egyetem, Filozófia Intézet

szekrenyes.istvan@btk.elte.hu

ORCID: [0009-0000-3503-8402](https://orcid.org/0009-0000-3503-8402)

Absztrakt

Az ITIdata (https://itidata.abtk.hu/wiki/Main_Page), a HUN-REN BTK Irodalomtudományi Intézet és a DigiPhil szemantikus adatbázisa az elmúlt három év során számos irodalomtudományos projekt metaadatainak adott helyet. Felhasználói oldalról érkeztek olyan igények, amelyek a beépített SPARQL lekérdező felület helyett könnyebben kezelhető keresők létrehozását inspirálták.

A Python nyelven implementált keresőalkalmazás célja egy egységes és rugalmas keresőmotor létrehozása volt, amely valamennyi ITIdata adatbázis számára könnyen kezelhető és testre szabható megoldást biztosít. Az alkalmazás lehetővé teszi a különböző adatbázis-specifikációkhoz való egyszerű adaptációt, beleértve a keresési opciók struktúráját, az eredménytáblák rendezési elveit és a lekérdezési módokat. Az alkalmazás egy JavaScripttel vezérelt webes interfészen keresztül érhető el, amely JSON konfigurációs fájlokat, SPARQL-lekérdezéseket, valamint szükség esetén MySQL-sémákat használ az egyes adatbázisokhoz való illeszkedéshez.

Tanulmányunkban három olyan kereső bemutatására vállalkozunk, amelyek mindegyike egy-egy irodalomtudományos projekt, *A magyar irodalomtörténet bibliográfiája*, a *Századfordulós sajtóbibliográfia*, valamint a *Krúdy Gyula-bibliográfia* köré szerveződik.

Kulcsszavak: bibliográfiai adatbázis, szemantikus adatbázis, keresőfelület, SPARQL, MySQL

¹ 2022-2.1.1-NL-2022-00009-es számú projekt a Kulturális és Innovációs Minisztérium Nemzeti Kutatási Fejlesztési és Innovációs Alapból nyújtott támogatásával, a 2022-2.1.1-NL pályázati program finanszírozásában valósult meg.

Abstract

New Search Interfaces for the ITIdata Semantic Database

ITIdata (https://itidata.abtk.hu/wiki/Main_Page), the semantic database of the HUN-REN BTK Institute for Literary Studies and DigiPhil, has hosted the metadata of numerous literary research projects over the past three years. User feedback indicated a need for more user-friendly search tools as an alternative to the built-in SPARQL query interface.

The goal of the search application, implemented in Python, was to develop a unified and flexible search engine that offers an easily manageable and customizable solution for all ITIdata databases. The application allows for straightforward adaptation to various database specifications, including the structure of search options, sorting principles for result tables, and query methods. The tool is accessible through a JavaScript-driven web interface, which uses JSON configuration files, SPARQL queries, and, where necessary, MySQL schemas to integrate with individual databases.

In this paper, we present three search engines, each developed for a specific literary research project: *Bibliography of Hungarian Literary History*, *Bibliography of Hungarian Belle Époque Press*, and the *Gyula Krúdy Bibliography*.

Keywords: bibliographic database, semantic database, search interface, SPARQL, MySQL

Az ITIdata a HUN-REN Bölcsészettudományi Kutatóközpont Irodalomtudományi Intézetének, valamint a DigiPhil csoportnak az adatbázisa. 2022-ben jött létre a Wikidata mintájára. A szolgáltatás az ELTE BTK TI Digitális Bölcsészet Tanszékén létrehozott ELTEdata² mintáját követi, az ott fejlesztett Wikibase-alapú szoftverre épül. Felépítésében is követi a Wikidata-t, az állítások (property), illetve a különböző projektek specifikációi tekintetében is. Az adatbázisban a DigiPhil és az Irodalomtudományi Intézet irodalomtudományos projektjei kapnak helyet, ideértve a digitális kritikai kiadásokhoz kapcsolódó bibliográfiai elemeket, valamint a különböző (irodalom)tudományos kutatások eredményeit, bibliográfiai rekordjait.³

A legfrissebb adatok szerint az adatbázisban 280 342 személynév rekord, 17 778 település rekordja, valamint 2590 foglalkozás, tisztség, rangnév rekordja található. Az ezen felül megtalálható bibliográfiai rekordokkal együtt összesen 490 700 rekordra lehet most keresni az Irodalomtudományi Intézet és a DigiPhil szemantikus tudásgráfiában. Az ITIdata-hoz –

² https://eltedata.elte-dh.hu/wiki/Main_Page

³ Lásd erről még, az egyes kutatások adatairól is: Dobás Kata – Fellegi Zsófia – Palkó Gábor, *A kis gömböc meséje – az ITIdata irodalomtudományos adatbázis fejlesztése 2022–2023-ban*, In: Új technológiákkal, új tartalmakkal a jövő digitális transzformációja felé: 32. Networkshop országos konferencia: 2023. április 12–14. Pannon Egyetem, Veszprém, HUNGARNET Egyesület, Budapest, 2022, 192–198.

a Wikidata-hoz hasonlóan – egy beépített keresőfelület tartozik, ahol SPARQL-lekérdezések segítségével lehet különböző kereséseket végrehajtani.⁴

A közel húsz irodalomtudományos projekt közül jelenleg háromhoz kapcsolódik saját fejlesztésű keresőfelület: *A Magyar irodalomtörténet bibliográfiájához* (MIB), a *Századfordulós sajtóbibliográfiához*, valamint a *Krúdy Gyula-bibliográfiához*.

A Magyar irodalomtörténet bibliográfiájának migrációjáról 2024-ben írtunk részletesen Tüskés Anna kollégámmal.⁵ Közel 110 000 rekord került át a migráció során egy korábbi felületről⁶ az ITIdata-ba. Az adatbázisban a bibliográfiai rekordokon túl található egy kiterjedt sajtóbibliográfia is, amely a bibliográfiában található, periodikákban fellelhető (tudományos) cikkek entitásalapú állításokhoz sorolása miatt jött létre; valamint kiadók, településnevek és az új szakrendhez szükséges elemek.⁷ Ez a legkiterjedtebb rekordszámú irodalomtudományos projekt az ITIdata-ban, szakmai vezetője Császtvay Tünde.⁸ Fontos kiemelni továbbá, hogy a másik két projekthez képest ez egy folyamatosan bővülő bibliográfia, a gyűjtési munkálatokat, a folyóiratok szemlélését az Irodalomtudományi Intézet Bibliográfiai osztályának munkatársai végzik.⁹

A Századfordulós sajtóbibliográfia elnevezésű projekt szakmai felelőse Wirágh András, az ő gyűjtésére alapozódott a sajtóbibliográfia.¹⁰ Nevéből következően a századfordulón létezett lapok rekordjait tartalmazza. Összesen 2280 sajtótermék került leírásra.

A Krúdy Gyula-bibliográfia Krúdy Gyula életében megjelent műveinek rekordjait tartalmazza – egyelőre az 1918-ig terjedő időszak szövegeit – amelyek kötetekben, periodikákban jelentek meg.¹¹ Összesen 5635 rekord található a bibliográfiában, a projekt szakmai felelőse szintén Wirágh András.

4 <https://query.itidata.abtk.hu/>

5 Dobás Kata – Tüskés Anna, *A magyar irodalomtörténet bibliográfiájának migrációja az ITIdata szemantikus adatbázisba. Gyakorlatok és tanulságok* In: Tick, József; Kokas, Károly; Holl, András (szerk.) *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, Budapest, Magyarország: Hungarnet Egyesület, 2024, 87–95.

6 <http://www.iti.mta.hu/mib/index.html>

7 Ez utóbbira néhány példa: tárgyszavak, temporális korszakolási fogalmak, történeti és földrajzi elhelyezkedés szerinti, illetve nyelvi alapú besorolások, elméleti és módszertani tudományos irányzatok, szemléleti elvek, vizsgálati módszerek stb.

8 A projekt nyitóoldala, ahol a keresőfelület linkje is megtalálható: https://itidata.abtk.hu/wiki/A_Magyar_Irodalomt%C3%B6rt%C3%A9net_bibliogr%C3%A1fi%C3%A1ja

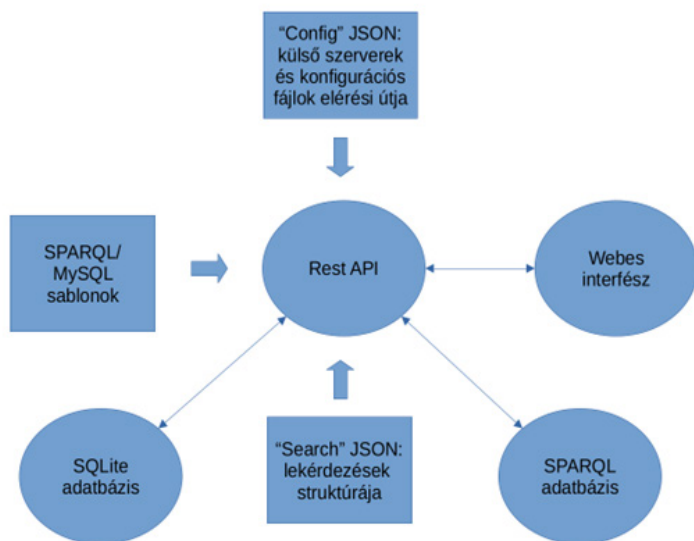
9 Erről lásd még részletesen: <https://kattatudomanyra.hu/az-irodalom-rejtett-halozatai/wikidata/>

10 Nyitóoldal keresőfelülettel: https://itidata.abtk.hu/wiki/Sz%C3%A1zadfordul%C3%B3s_sajt%C3%B3bibliogr%C3%A1fia

11 Nyitóoldal keresőfelület linkjével: https://itidata.abtk.hu/wiki/Kr%C3%BAdy_Gyula-bibliogr%C3%A1fia

Miért volt szükség új keresőfelületekre, amikor rendelkezésre állt a SPARQL lekérdező-felület? Az ITIdata felhasználóinak köre elsősorban irodalomtörténészekből, bölcész kutatókból áll. Elsősorban erről az oldalról érkeztek hozzánk olyan kérések, amelyek ezt indokolták. A SPARQL lekérdezőnyelv nem volt mindenkinek komfortos, hiába volt hozzá beépített lekérdezősegéd. Az egyik lehetőség az lett volna, hogy példalekérdezésekkel segítjük elő a különböző projektek nyitóoldalain a kereséseket, de végül új felületek létrehozása mellett döntöttünk. A cél az volt, hogy egy minden projekt számára könnyen adaptálható felületet hozzunk létre, biztosítva az átjárást is az egyes irodalomtudományos kutatások eredményei között.

Technikai oldalról ennek megvalósításához egy egyszerűen konfigurálható keretrendszer létrehozására volt szükség, ami a felhasználók részére egy webes interfészen HTML-űrlapok segítségével biztosítja a keresési paraméterek megadását, majd az automatikusan generált SPARQL-lekérdezések eredményének átlátható formában történő megjelenítését. A keretrendszer központi alapját egy Python nyelven implementált REST API jelenti, ami a keresőfelület testreszabásához szükséges információkat JSON konfigurációs állományokból olvassa ki, a lekérdezéseket pedig az interfészen megadott paraméterek, illetve SPARQL-sablonok segítségével állítja össze és továbbítja a SPARQL-végpont felé. Néhány speciálisabb projekt esetében, entitásazonosítókat tartalmazó, illetve az eredeti Wikibase adatbázisból migrált SQLite-adatbázisokkal is kommunikál. A kapott eredménytáblákat a JSON konfigurációs fájlok beállításai szerint dolgozza fel, végül továbbítja őket az interfész felé. Maga a megjelenítés Javascriptek és CSS-stíluslapok segítségével történik. A keretrendszer moduljai közötti kapcsolatot az 1. ábra szemlélteti.



1. ábra: A keretrendszer felépítése

A legalapvetőbb beállításokat (végpontok URL-címe, SPARQL-sablonok és a további konfigurációs fájlok elérési útja) a *config.json* nevű állomány tartalmazza, illetve egy külön JSON-fájlban található az egyes projektek specifikációinak leírásai. Utóbbi akár többféle, az interfészen választható kereséstípust is (pl. keresés személyekre, helyekre, kiadványokra valamelyik gyűjteményben) tartalmazhat a *search* attribútum listaelemként.

Az alábbi példa a Krúdy Gyula-bibliográfia specifikációján szemlélteti a JSON konfigurációs állomány szerkezetének egy részletét:

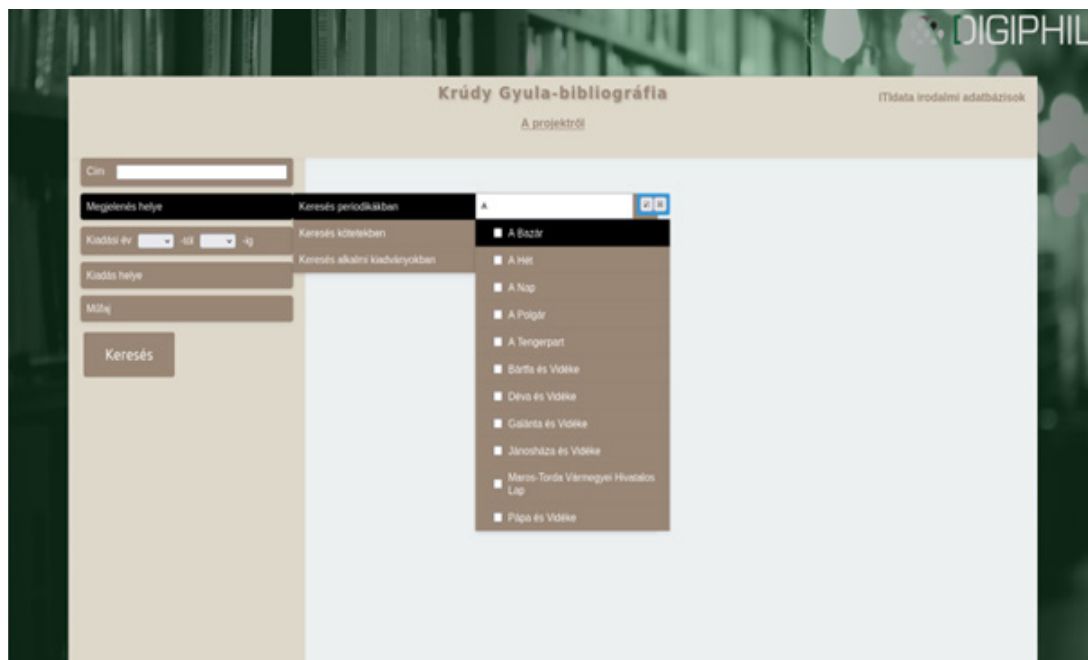
```
{
  „title”:”Krúdy Gyula-bibliográfia”,
  „link”:”https://itidata.abtk.hu/wiki/Kr%C3%BAdy_Gyula-bibliogr%C3%A1fia”,
  „search”: [
    {
      „id”:”search1”,
      „label”:”Kiadványok keresése”,
      „type”:”sparql”,
      „params”: {
        „itemLabel”:”,
        „yearFrom”:”1892”,
        „yearTo”:”1918”,
        „placeOfPublicationLabel”:[],
        „publishedINLabel”:[],
        „genreLabel”:[]
      },
      „defaultOrder”:”itemLabel”,
      „sparqlTemplate”:”q1”,
      „results”:{
        „format”:”table”,
        „items”: [
          {
            „type”:”url”,
            „label”:”Cím”,
            „src”:”item”,
            „text”:”itemLabel”
          }
        ]
      }
    }
  ]
}
```

Az egyes kereséstípusok különféle tulajdonságai közül (azonosító, a weboldalon megjelenítendő címke, a lekérdezés típusa/nyelve stb.) a *params* attribútum sorolja fel a felhasználó által megadható paramétereket, többségüket üres értékkel, de alapértelmezés (mint pl. a kiadvány megjelenési évének tartománya) beállítására is van lehetőség. A *results* attribútum tartalmazza az eredmények megjelenítéséhez szükséges információkat, a táblázat

egyos oszlopait és a bennük szereplő értékek jellemzőit (hiperhivatkozás vagy szimpla szöveg), illetve hogy melyik paraméterből olvashatók ki. A *defaultOrder* a sorok rendezésének alapértelmezett alapját, a *sparqlTemplate* tulajdonság pedig a kereséshez tartozó SPARQL-sablon azonosítóját adja meg.

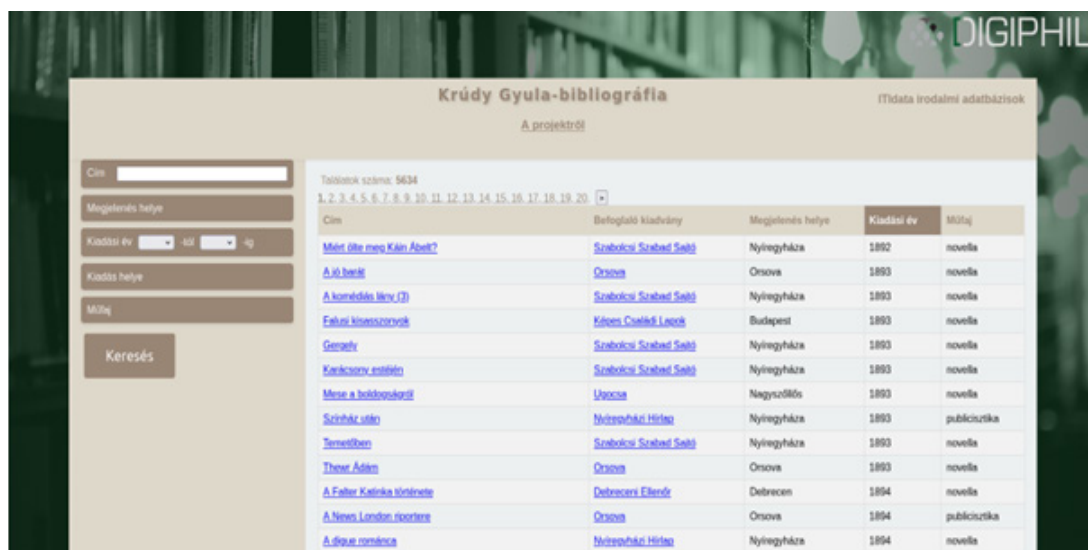
A keresési paraméterek beállítására szolgáló HTML-űrlapok struktúráját a *menuStructure* attribútum írja le. Mint az alább következő példában látható, többszintű menürendszerek létrehozására is van lehetőség. A példában az *itemLabel* paraméternél szimpla szövegként megadhatjuk a keresett kiadvány címét, vagy annak egy részletét, a „Megjelenés helye” (PublishedINLabel paraméter) esetén pedig az interfészen megnyitható egy almenü, ahol periodikák, önálló kötetek, illetve alkalmi kiadványok közül választhatunk. Ezeket nem szükséges a JSON-fájlban felsorolni, helyette a *subquery* attribútummal meghívhatunk egy SPARQL-lekérdezést, ami ebben az esetben a lehetséges megjelenítési helyek listájával fog visszatérni. Ha a *subsearch* tulajdonság értéke *true*, akkor a webes felületen kereshetünk is közöttük (lásd 2. ábra).

```
„menuStructure”: [
  {
    „optionName”:”Cím”,
    „param”:”itemLabel”,
    „inputType”:”text”
  },
  {
    „optionName”:”Megjelenés helye”,
    „subOptions”: {
      „class”:”submenu”,
      „optionList”: [
        {
          „optionName”:”Keresés periodikákban”,
          „subOptions”: {
            „class”:”subsubmenu”,
            „subSearch”:true,
            „param”:”publishedINLabel”,
            „startID”:1,
            „type”:”sparql”,
            „subquery”:”periodic”
```



2. ábra: Keresés a beágyazott lekérdezés eredményei között

A beállított limitet meghaladó számosságú találat esetén az eredménytáblát több oldalon lapozhatjuk, a sorokat az oszlopnevekre kattintva van lehetőség növekvő vagy csökkenő sorrendbe ártrendezni. A 3. ábrán a kiadási év szerint növekvő sorba rendezett találatok első oldalát láthatjuk.



3. ábra: A találatok megjelenítése és rendezése

További terveink között szerepel a webes interfész funkciónak bővítése (keresési eredmények különböző fájlformátumokban történő mentése, kiválasztott szűrők megjelenítése és törlése stb.), az SQLite-adatbázisokba történő migráció és frissítés automatizálása, adatvizualizációs technológiák integrálása.

Összegzésként elmondható, hogy az ITIdata-keretrendszerre épülő, Python-alapú keresőmotor sikeresen hidalja át a SPARQL-felület meredek tanulási görbáját, miközben megőrzi a szemantikus adatbázisok rugalmasságát. A bemutatott példákon keresztül igazoltuk, hogy a JSON-konfigurációkon alapuló REST API és a JavaScript-vezérelt webes interfész minimális fejlesztői beavatkozással adaptálható eltérő projektigényekhez. Eredményeink azt mutatják, hogy az új felület nemcsak gyorsabb és intuitívabb keresést tesz lehetővé az irodalomtörténészek számára, hanem elősegíti az egyes adatbázisok közötti átjárhatóságot is. A jövőben az automatikus SQLite-migráció, a valós idejű frissítés és az interaktív adatvizualizáció integrálása tovább erősítheti az ITIdata szerepét a hazai digitális bölcsészet infrastruktúrájában.

Bibliográfia

Dobás Kata – Fellegi Zsófia – Palkó Gábor, *A kis gömböc meséje – az ITIdata irodalomtudományos adatbázis fejlesztése 2022–2023-ban*, In: Új technológiákkal, új tartalmakkal a jövő digitális transzformációja felé: 32. Networkshop országos konferencia: 2023. április 12–14. Pannon Egyetem, Veszprém, HUNGARNET Egyesület, Budapest, 2022, 192–198.

Dobás Kata – Tüskés Anna, *A magyar irodalomtörténet bibliográfiájának migrációja az ITIdata szemantikus adatbázisba. Gyakorlatok és tanulságok* In: Tick, József; Kokas, Károly; Holl, András (szerk.) *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, Budapest, Magyarország: Hungarnet Egyesület, 2024, 87–95.

A keresőfelületek linkjei:

<https://mibkereso.iti.abtk.hu/>

<https://sajtokereso.iti.abtk.hu/>

<https://krudykereso.iti.abtk.hu/>