Networkshop 2008, Dunaújváros

Fejlett kereső és lekérdező eszközök egy elektronikus szakfolyóirathoz (IBVS)

Holl, A., holl @ konkoly . hu
Erdődi, P., don @ konkoly . hu
MTA Konkoly-Thege Miklós Csillagászati Kutatóintézete

Az előadás vetített anyaga:
http://www.konkoly.hu/staff/holl/Duna/dunausli.pdf

Tartalmi kivonat:

Az Information Bulletin on Variable Stars egy kis, elektronikus formában is megjelenő csillagászati szakfolyóirat (lásd Holl, Networkshop 2001). A folyóiratban publikált cikkekben, ábrák között, valamint a cikkekhez tartozó adatállományokban való keresés céljából új eszközt fejlesztettünk. A kereső jellemző tulajdonsága, hogy a megszokott, hasonló programoktól elvárható funkciókon túl (keresés a meta-adatokban és a teljes szövegben) keresni tud ábrákat és adatállományokat is. Továbbá lehetőség van csillagászati objektumok keresésére, objektumnevek helyett: ugyanazon objektum szerepelhet különböző neveken, és bármelyik névvel megtalálható. Hasonló a szerzők nevére való keresés is: a többféle változatban használt nevek bármelyik írásmódja használható a keresési feltételben. Ezeket a tulajdonságokat nem csupán lokális szótárakkal, de GRID- avagy Virtuális Obszervatórium-jellegű funkcionalitással: a hálózaton elérhető szolgáltatások segítségül hívásával érjük el. A fenti rendszerre alapozva olyan lekérdezési lehetőséget is készítünk, melynek segítségével objektumnevek és adattípusok (pl. adott típusú ábra) megadásával lehet majd külső adatbázisokból linkeket generálni az IBVS-ben közölt információkra.

Advanced Search Tool and DataService for an electronic journal: the IBVS

Abstract:

The Information Bulletin on Variable Stars is a small electronic journal in astronomy (see: Holl, Networkshop 2001). We have developed an advanced search tool which can be used on the full text, meta-data, figures and electronic data tables. It is capable of searching for "objects" as opposed to "object names", and "authors" instead of "author names".
These features are achieved with the use of local dictionaries, and with the invocation of other services on the Internet: a true GRID- or Virtual Observatory-functionality. Based on this tool an associative search tool is being built too: with that data in IBVS could be linked from external databases using object names.

A folyóirat

Az Information Bulletin on Variable Stars (IBVS) egy kis szakfolyóirat, mely rövid cikkeket közöl

a változócsillagászat témakörében. Az IBVS a Nemzetközi Csillagászati Unió 27-es és 42-es Kommisszióinak lapja, amelyet az MTA Konkoly-Thege Miklós Csillagászati Kutatóintézete ad ki (Holl A., Networkshop 2001).

Miért kell új kereső?

A folyóirat elektronikus változatának megindulása óta (1994) a tárolt anyag mennyisége megnövekedett - digitalizálásra kerültek a cikkek egészen az 1961-ben megjelent első számig, és több, mint tíz év már eredetileg is számítógépes szövegszerkesztéssel készült anyaga került a webre. A 90-es évek közepén az elektronikus IBVS-hez készült keresőrendszer mára elavult. Új keresőre lett szükség, amely kielégíti a következő igényeket: i.) fejlettebb logikai feltétel és reguláris kifejezés kezeléssel rendelkezik; ii) bővebbek a meta-adatokban való keresés lehetőségei (pl. objektumnév, tartalmi kivonat szövegben szereplő kifejezések); iii) segíti a tartalmi feltárást: keresni lehet ábrákat vagy adatállományokat is; iv.) képes különböző nevek szerint is megtalálni ugyanazt az objektumot, illetve a szerzők neveinek különböző írásmódja sem jelent akadályt az azonosításukban; v.) szebben, informatívabban jeleníti meg a keresés eredményeit.

A Google megjelenése magasabbra állította a mércét: egy modern keresőrendszernek funkcionalitását, prezentációját és sebességét tekintve is állnia kell a versenyt. Megkérdezhetjük: miért nem használjuk akkor magát a Google-t? Azért nem, mert az IBVS túlságosan speciális. Egyrészt a publikus keresők felől érkező találatok - a kezdeti időkben, amikor ezeket a keresőket még beengedtük az IBVS anyagába - túlnyomó részt tévesek voltak, másrészt ezek a keresők nem rendelkeznek elegendő ismerettel a folyóirat szerkezetéről. Még a Google Images sem talál meg minden képet - az IBVS esetében csak saját keresővel tudtuk megoldani, hogy egy adott csillagra vonatkozó, adott típusú ábrát tökéletes biztonsággal meg tudjon találni az olvasó. A Google-nél jobb keresési lehetőséget biztosít egy speciális csillagászati szakirodalmi kereső rendszer: a NASA Astrophysics Data System Abstract Service - de még ez sem elégítette ki az igényeinket. Ez a rendszer sem ismeri a csak erre a folyóiratra jellemző tulajdonságokat, és heti rendszerességű frissítése sem elegendő: aki kifejezetten az adott folyóiratban akar keresni, az meg akarja találni a fél órával korábban közölt cikkeket is.

Tartalmi feltárás

Az olvasó (ebben az esetben a csillagász) többnyire nem a bibliográfiai adatok alapján keres a szakirodalomban. Nagy szerepe van a tematikus csoportosításnak, kulcsszavaknak, s különösen a csillagászati objektumok (az IBVS esetében többnyire változócsillagok) neveinek. Sokszor arra van igény, hogy egy adott csillagról adott típusú információt: ábra-típust (keresőtérkép, fénygörbe) vagy adatállományt találjanak meg. Az IBVS esetében a szerzők szabadon adhatnak meg kulcsszavakat, és ez sajnos, nem tesz lehetővé hatékony keresést. A cikkekben szereplő objektumok nevének megjelölése, ellenőrzése megtörtént: ezt az információ felhasználható az új keresőprogramban. Ugyancsak feldolgoztunk minden ábrát és a cikkekhez kapcsolódó adatállományt, itt is megjelölve

az objektumokat, és - ebben az esetben már kötött készletből választható - kulcsszavakat adtunk meg. Felismertük, hogy a folyóirat sok évtizedes története során bizonyos típusú adatokból olyan mennyiséget halmozott fel, hogy ezek az eredeti cikkekből kiragadva, adatbázisként is használhatók - bár el kell ismernünk, meglehetősen inhomogén adatbázisként. A felhasználó sok esetben nem is kíváncsi az egész cikkre, hanem csupán, mondjuk, egy benne közölt ábrára. Ezért lehetővé tettük, hogy minden ábrához, adatállományhoz közvetlenül hozzá lehessen férni egyedi azonosítója alapján. (Természetesen az ábra vagy adatállomány mellett szerepelnek annak a cikknek a tömör bibliográfiai adatai, amelyben eredetileg közlésre került. Ugyancsak megjeleníthetők az adott cikkelemre vonatkozó meta-adatok, és lehetőség van különböző letöltési formátumok közül választani.)

Összefonódó, humán- és automata felhasználóknak nyújtott információs szolgáltatások

A csillagászatban - haszon-talan tudomány lévén - rengeteg szabadon hozzáférhető információ áll a kutatók rendelkezésére, és a különböző információs szolgáltatások erősen összefonódnak. Az IBVS egyes cikkeihez el lehet jutni a legfontosabb csillagászati információszolgáltatóktól: a már említett bibliográfiai rendszerből, az ADS-ből, vagy a strasbourgi CDS Simbad adatbázisból, és az IBVS HTML változatában is rengeteg hiperhivatkozás található. A sok hiperhivatkozás követése néha már megnehezíti az informálódást: felmerül az igény arra, hogy egyetlen oldalon, tömören prezentálják a különböző forrásokból származó információt. Ezért az új keresőprogramra építve olyan szolgáltatás létrehozását tervezzük - ez lesz az IBVS DataService - ami immár nem egyedi azonosítók, hanem a csillagászati objektum neve, és a cikkelemtípus alapján címezhető. Az IBVS DataService nem csak közvetlenül a felhasználó, hanem más szolgáltatások számára is tud majd adatokat szolgáltatni, melyek alkalmasak lesznek arra, hogy pl. egy weblapba beépüljenek. Felmerült az az igény is, hogy az ADS számára csatlakozási felületet (API) adjunk, teljes szövegű keresési feladatok továbbítására.

Névfeloldás

A kereső készítése során meg kellett oldani egy problémát: a csillagászati objektumoknak több - sokszor akár több tucatnyi - neve is lehet. A meta-adatok karbantartása során törekedhetünk arra, hogy lehetőleg egy bizonyos katalógusban használt neveket használjunk, vagy lecseréljük ezekre a nevekre a szerzők által használt másfajta neveket - de azt már nem szabhatjuk meg, hogy a felhasználó milyen néven keressen egy csillagot. Az egyes objektumok neveinek összegyűjtése és karbantartása meghaladja a lehetőségeinket - és különben is, van egy adatbázis, a Simbad, amelyik már régóta sikeresen végzi ezt a feladatot! Miért ne használhatnánk az IBVS keresőrendszeréhez ezt a külső tudásbázist? Ugyanerre a célra felhasználható a moszkvai GCVS szolgáltatás is. A Simbad és a GCVS ez irányú információi - bár nagyrészt megegyeznek - mégsem teljesen fedik egymást. A jelenlegi megvalósításban a felhasználó megjelölheti, akar-e névfeloldást használni, és ha igen, akkor (szegényes tartalmú) lokális szótárat, a Simbad-ot vagy a GCVS-t részesíti előnyben.

Az objektumnevek után már könnyen adódik az ötlet, hogy hasonló névfeloldás alkalmazható a

szerzők nevei esetében is. Itt az ADS az a szolgáltató, aki rendelkezik ilyen információval. A keresőben választható ez esetben is a lokális szótár és az ADS használata, illetve el lehet tekinteni a névfeloldástól.

Külső szolgáltató használata esetén a keresőprogram HTTP protokoll szerint lekérdezi az adott név változatait, és feldolgozza a kapott HTML, XML vagy ASCII szöveges állományt. Harmadik fél által nyújtott hálózati szolgáltatások igénybevétele a GRID illetve a Virtuális Obszervatórium kezdeményezések törekvéseivel rokonítja az új IBVS keresőprogramot.

Megvalósítás

A szerzők egyike (E.P.) az ELTE-IK programtervező szak kooperatív képzésén vett részt az MTA KTM CsKI-ben, ez adott alkalmat a keresőprogram megírására. A fejlesztés Perl nyelven történt. A nyelv használata mellett több érv szólt, ezek közül a legfontosabbak:

- az interpreter teljesen szabad szoftver,
- a nyelv sajátossága a reguláris kifejezések igen magas szintű támogatottsága – amely egy keresőprogramban elengedhetetlen,
- igen jó az weben keresztüli támogatottság: dokumentációk, segédletek, példák és hasznos kiegészítő csomagok nagy számban fellelhetők a világhálón.

A fejlesztés alatt figyelembe kellett vennünk az CsKI hardver és szoftver lehetőségeit/korlátait. A ma már hagyományosan bevett, adatbázis alapú keresőmódszerek alkalmazására nem volt lehetőség, így indexfájlokat alkalmaztunk adatbázis táblák helyett.
A módszer hátrányai:

- az adatbázis rendszerek intelligens memóriakezelése/puffer technikái lényegesen javítják a keresés sebességét (megjegyzendő, hogy tervben van egy hatékony puffer technika implementálása a jelenlegi rendszerhez);
- a kereső/lekérdező program igen speciális, nehézkesebb az új kereső lekérdezések implementálása, szemben az AB alapú rendszerekkel, ahol ez csak egy új SQL parancs megírását jelenti.

A módszer előnyei:

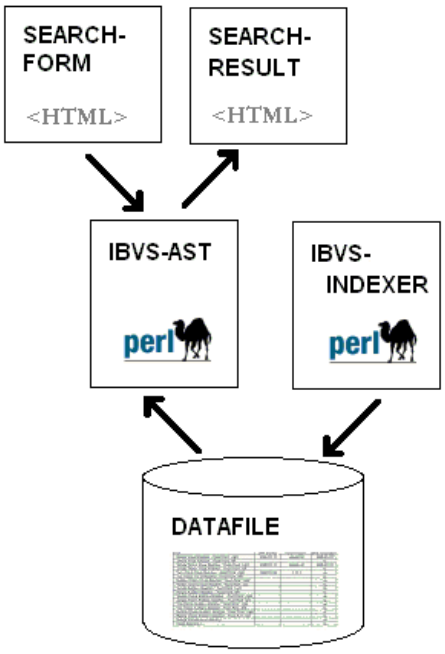
- az egyik előny éppen ez utóbbi hátrányban rejlik: az adatfájl és a keresőszkript között igen erős a koherencia, amely meggyorsítja a keresést: a program teljesen az adatfájlhoz van igazítva;
- a kliens és az adatfájlok közé nem integrálódik egy AB rendszer, amelynek általánosságából mindenképpen következik, hogy „felesleges” szolgáltatásokat tartalmaz;
- nemcsak a program követi jól az adatfájl felépítését, hanem az adatfájl struktúrája is a keresést támogatja.

Összefoglalva: a hardver és szoftver lehetőségek számbavételével és a keresések gyakoriságának figyelembevételével - mivel igen speciális, nem széles körben használt tartalomról van szó -, a jelenlegi keresőrendszer előnyösebb az Intézet számára, mint egy „korszerűbb”, AB alapú rendszer.

A kereső tervezésénél fontos szempont volt, hogy a kezelőfelület egyszerű és funkcionális legyen,

valamint hogy a program URL-ben is teljes paraméterezési lehetőséggel meghívható legyen.

Az alkalmazás felépítése:



Felhasználói réteg: a **kereső űrlap** és a **találatok** megjelenítése

Az alkalmazás lényegi része a két modul: az **indexelő** és a **kereső**. Az előbbi hozza létre/frissíti az adatfájlt, a második pedig a keresést végzi.

A strukturált **adathalmaz.**

A kereső szkript elérése: `http://www.konkoly.hu/cgi-bin/advanced_search.pl`

Paraméterezés (egyik sem kötelező paraméter!):

1. Keresőfeltételek

```
CaseSensitiveCheckbox=[on|off]
IntervalBottom=IBVS szám
IntervalTop=IBVS szám
ObjectNameResolution=[none|Local|GCVS|Simbad]
AuthorNameResolution=[none|Local|ADS]
Text=szöveg
ExactPhraseTextCheckbox=[on|off]
Title=szöveg
AuthorList=szöveg
AbstractText=szöveg
AObjectName=objektumnév
PublicationDateFromMonth=hónap száma
PublicationDateFromYear=év
PublicationDateToMonth=hónap száma
PublicationDateToYear=év
PaperOrFigure=[Paper+with+figure|Figure+only]
IBVSfig=szöveg
FObjectName=objektumnév
IBVSfigKey=ábra típus
PaperOrDatafile=[Paper+with+datafile|Datafile+only]
DObjectName
DataKey=ábra típus
```

2. A kimenet szabályozása

```
ShowSearch=[on|off]
ShowResult=[on|off]
ShowHeader=[on|off]
ShowFoot=[on|off]
ShowTotal=[on|off]
ShowTime=[on|off]
```

A keresőprogram URL-ben való meghívhatósága lehetővé teszi a más alkalmazásokba való beépítését, ugyanakkor módot adott az intenzív, batch-módban való tesztelésre. A kapott teljesítmény elfogadható: a keresési idő néhány, vagy néhányszor tíz másodperc a legtöbb feladat esetében.

A dokumentáció a program forráskódba ágyazva készül.

A keresőt hamarosan hozzáférhetővé tesszük az IBVS felhasználói számára, és meg fog lenni az "IBVSlatest" szolgáltatásban is: ez a legfrissebb IBVS számokat jeleníti meg az olvasóknak.

Irodalom / URL-ek

Holl A., 2001, "Elektronikus folyóiratok a természettudományok területén - egy hazai példa", Networkshop előadás http://www.konkoly.hu/staff/holl/sopron/sopron.html

NASA Astrophysics Data System FAQ: http://doc.adsabs.harvard.edu/abs_doc/faq.html

CDS Simbad adatbázis: http://simbad.u-strasbg.fr/simbad/

General Catalogue of Variable Stars: http://www.sai.msu.su/groups/cluster/gcvs/gcvs/