PHYSHUN elosztott alapú kersőrendszer

Szalay Istvánné <szalay@sunserv.kfki.hu>
KFKI RMKI Számítógép Hálózati Központ
Kadlecsik József <kadlec@sunserv.kfki.hu>
KFKI RMKI Számítógép Hálózati Központ
Köveshegyi László <koeves@sunserv.kfki.hu>
KFKI RMKI Számítógép Hálózati Központ

Kivonat

A PhysHun keresőrendszer a magyarországi fizikus szakmai közösségek lokálisan felhalmozódott, digitális formában létező dokumentumainak kereshetőségét biztosítja magyar és angol nyelven, és egyúttal részét képezi a világméretű PhysNet rendszernek.

A PhysHun keresőrendszer az ITEM-2003 pályázat keretében készült az IHM támogatásával.

A PhysHun bővíthető rendszer, célja a magyarországi fizika tárgyú web-dokumentumok kereshetővé tétele.

A rendszer a MySQL és a mnoGoSearch szabad szoftverek felhasználásával működik, a lokális web-adatbázisokra és Dublin Core szabvány szerint előállított meta információkra épül.

A rendszer jelenleg hat intézmény dokumentumait foglalja magában. Az adatbázisok együtt és külön-külön is kereshetők.

Bevezetés

A projekt célja a fizika érdeklődési körök információs igényét kielégítő kapcsolati rendszer megvalósítása, amely a fizika iránti érdeklődés látókörébe eső hazai és nemzetközi, többnyelvű információforrások intelligens kereshetőségét biztosítja.

Az ország számos intézményében, a fizikával foglalkozó kutatóhelyeken és egyetemek fizika tanszékein jelentős mennyiségű fizikával kapcsolatos, kutatási eredményeket, vagy oktatási anyagokat tartalmazó, magyar, angol, esetenként más nyelvű információ, dokumentum halmozódott fel a webszervereken. A fizikusok, egyetemi hallgatók, fizikatanárok, középiskolások számára azonban jelenleg nem áll rendelkezésre olyan megfelelő keresőeszköz, amellyel az országban működő intézmények web-oldalait gyorsan áttekinthetnék, azokban gyors kereséssel tájékozódhatnának, és a számukra fontos információforrásokhoz hozzáférhetnének. Sok esetben nem tudnak arról gyorsan információt szerezni, hogy másik, hazai vagy külföldi, hasonló profilú intézmény honlapján milyen, számukra is értékes anyagok vannak.

Az általános keresőkben feltett keresőkérdésre rendszerint túl nagy számú válasz érkezik, amelyek egy része értéktelen, másik része külföldi szerverre mutat, ilyen módon nem biztosítják az értékes hazai információkhoz való hozzáférést.

Mi az oka a fenti problémának?

- A fő gond az, hogy a felhasználó kénytelen általános keresőt használni, mivel Magyarországon nem létezik egy, a fizika érdeklődési kör információs igényét kielégítő keresőrendszer.
- A szerverek többségén nem működtetnek keresőt saját oldalaik indexelésére.
- A magyar szervereken lévő, fizikával kapcsolatos dokumentumok minősége rendkívül vegyes. Többségük tartalmilag értékes ugyan, de az igénytelen, pontatlan kódolás miatt értékéből sokat veszítve a keresők számára nem eléggé használható
- A dokumentumok nincsenek kiegészítve meta-adatokkal, ezek hiánya pedig sok esetben "megtalálhatatlanná", vagy nehezen megtalálhatóvá teszik a web-en lévő anyagot, de legalábbis sokszor a releváns találatok a lista végére kerülnek. Ennek pedig az az oka, hogy a rangsorolási (ranking) funkcióval rendelkező kereső szoftver a kereséshez megadott kulcsszavak előfordulási helye és gyakorisága alapján megpróbálja a legjobb találatokat előbbre sorolni a találati listában.

A probléma elhanyagolásának egyik legkézzelfoghatóbb következménye lehet, hogy a diákok amúgyis megkérdőjelezhető fizika iránti érdeklődése még alacsonyabb szintre süllyed. A még érdeklődő tanulók is elkedvetlenedhetnek, ha a manapság egyre természetesebbé váló eszköz segítségével sem juthatnak hozzá olyan érdekes fizikai információkhoz, amelyek az egyre zsugorodó tananyagból kimaradtak. A kutatók és egyetemi hallgatók számára a jelenlegi helyzetben a fizikára vonatkozó hazai információforrások nem elég jól, nem elég hatékonyan férhetők hozzá. Az információs társadalom építésének korszakában feltétlenül szükséges, hogy az információs társadalom egyik alapját jelentő fizikával kapcsolatos ismeretek is könnyen kereshetők legyenek.

A hazai szükségleten kívül jogos igény merül fel a külföldi fizikus társadalom részéről is a jó hírnévvel rendelkező magyar intézmények web-en lévő anyagainak, kutatási eredményeket ismertető beszámolóinak és oktatási anyagainak megtalálása és eredményeik elektronikus publikálása iránt.

Dublin Core metaadatok

A korábbi elmélet és elképzelés, miszerint a meta elemek alkalmazásával a nagy általános keresőrendszereknek adunk információkat, egyre inkább háttérbe szorul. Hangsúlyt kapott viszont szerepük a speciális, testre szabott keresőket alkalmazó rendszerek (pl. PhysNet) és szakmai portálok (pl. SZEZAM) terén, valamint alapvető fontosságú együttműködő partnerszervezetek informatikai rendszerei közötti adatcsere-kapcsolathoz (pl. NDA = Nemzeti Digitális Adattár). A közös adatcsere protokoll alapja az Open Archives Initiative (OAI) Protocol for Metadata Harvesting (PHM) ajánlása, amely a Dublin Core (DC) Metadata Initiative szabványra épít. Reményeink szerint valamikor majd a PhysHun is egy ilyen rendszer része lesz.

A Dublin Core metaadatok generálását saját fejlesztésű metageneráló scripttel végeztük. A PhysHun Meta Maker kötelezően és nem kötelezően megadandó metaadatokat generál, s ezek egy részét - Title, Creator, Subject, Description- konvertálja egyszerű name ill. http-equiv típusú metatag-gé. - title, author, keywords, type, language, description - azon keresők részére, amelyek mégis figyelembe vesznek bizonyos metaadatokat. Alkalmazásával választható, hogy a metaadatokat beépítsük-e a dokumentum header részébe is, vagy csak külső metafile keletkezzék. A "beépített" metaadatoknak az az előnye, hogy egyes browserek a View/Page info-re kattintva megmutatják ezeket. A külső metafile generálás viszont a távlati elképzelések szempontjából (OAI, NDA) és az egységesség és átláthatóság szempontjából kívánatos.

A pályázatban foglalt célkitűzésre tekintettel a többnyelvűség megvalósítására a címet

legalább két nyelven adjuk meg. Magyar nyelvű dokumentum esetén magyarul és legalább egy idegen nyelven (a továbbiakban az egyszerűség kedvéért tételezzük fel, hogy angolul), idegen nyelvű dokumentumoknál pedig lefordítjuk és megadjuk a címet magyarul. A cím többnyelvű megadása kötelező, de lehetőleg a kulcsszavakat és a description-t is megadjuk több nyelven!

A DC.Subject metaelemnek kiemelt jelentőséget tulajdonítva többféle dokumentumot különböztetünk meg:

- Ha a dokumentum egy konkrét szakmai anyag (pl. könyv, cikk, oktatási anyag, szakdolgozat, egy bizonyos témájú konferencia, stb.), amelynek ismert a szerzője (konferencia esetén a szervezője) akkor megadjuk a PACS (Physics and Astronomy Classification Scheme) szerinti osztályozás kódját és a kódhoz tartozó szöveget (), magyar nyelvű dokumentum esetén az Ortelius Thesaurus angol-magyar nyelvű osztályozás szerinti tárgyszót és négy-öt kulcsszót legalább két nyelven. ().
  Az Ortelius Thesaurus (http://www.info.omikk.bme.hu/nkr1/CERIF/orteliustop.htm) szerepel a "Controlled vocabularies, thesauri and classification systems available in the WWW. DC Subject" című dokumentumban.
- Ha a dokumentum típusa más jellegű (pl. intézményi honlap, eseménynaptár, publikációs lista, stb.), akkor legjobb tudásunk szerint adjunk néhány, a dokumentum tartalmát jellemző kulcsszót legalább két nyelven.

Hangsúlyos szerepe van még a DC.Type metaelemnek, amellyel a dokumentum típusát, műfaját adjuk meg. A metaelem indexelése révén a felhasználónak módjában áll kiválasztással meghatározni, hogy milyen műfajú dokumentumok között akar keresni.

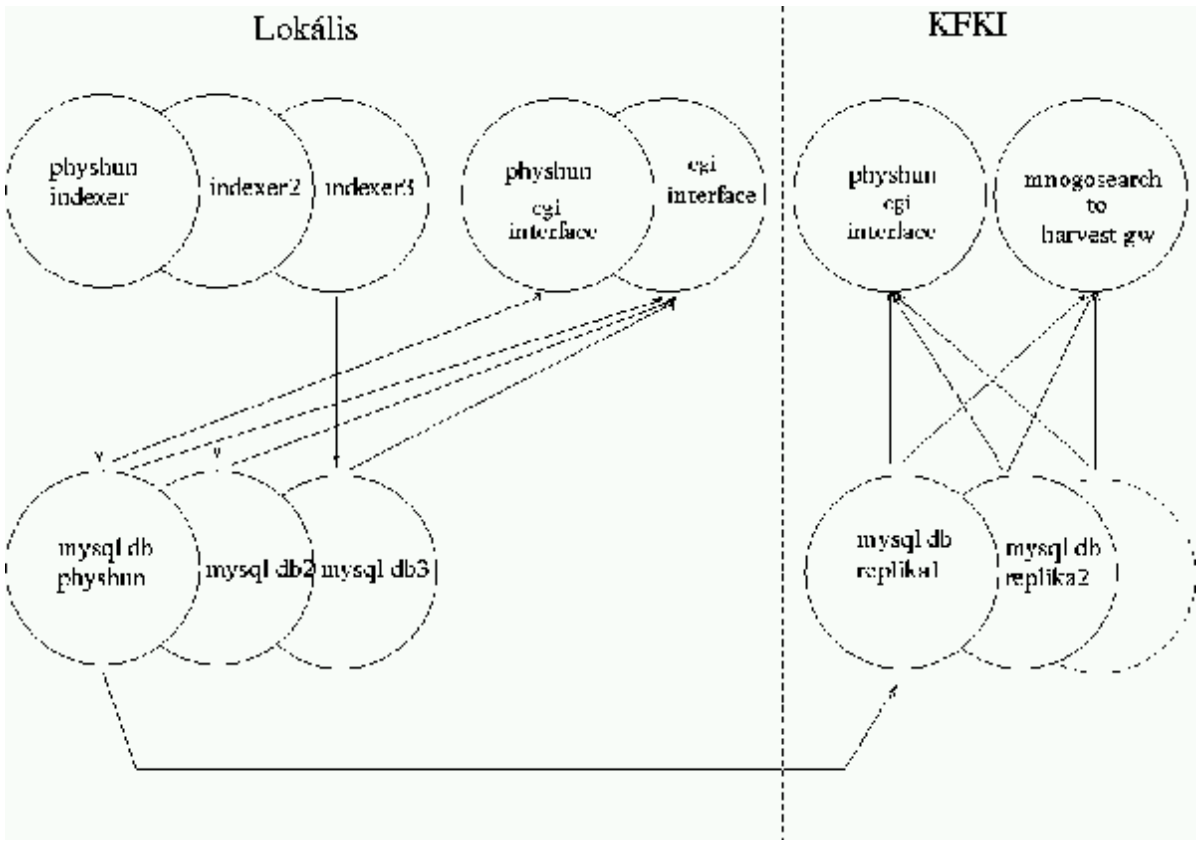
Ugyancsak a metainformációk alapján a keresés találati listájában jelezzük a dokumentum nyelvét

A PhysHun rendszer sémája

A PhysHunrendszer alapelemei:

- lokális mnoGoSearch indexerek
- lokális cgi interfészek
- lokális MySQL adatbázisok
- központi cgi interfész
- replikált MySQLadatbázisok
- mnoGoSearch to Harvest gateway a PhysNet integrációhoz

A fenti elemek közötti összefüggést az alábbi ábra mutatja:



Mysql replikáció a PhysHun projektben

A konzorciumi partnerek saját gépeiken indexelt adatai helyi mysql adatbázisba kerülnek, amely helyi adatbázisok szerverei master típusúak.
Ezeket a helyi adatbázisokat kell a központi physhun gépen slave mysql adatbázisokba tárolni, ennek eszköze a replikáció.
A master-slave viszonyt azért definiáltuk olyan módon, hogy a helyi adatbázisok legyenek a master típusúak, hogy egyetlen gép esetleges kiesése ne okozzon kiterjedt adatvesztést. Így a központi gépen slave típusú mysqld kell fusson.

A mysql replikáció logikája szerint egy slave mysql szerverhez egy és csakis egy master szerver tartozhat, ezért a központi gépen annyi példányban kell fusson slave típusú mysql szerver, ahány helyi adatbázist replikálni kell.
Emiatt az egyes slave mysql szervereket egymás port-jától, és az alapértelmezett mysql port-tól is eltérő port-okon kell futtatni. Ugyanez igaz a socket típusú hozzáférésre.
Ezeket a paramétereket slave mysql szerverenként különálló konfigurációs file-okban kell definiálni, ráadásul hozzáférési módonként (különféle kliensek számára) külön szekcióban.

A replikáció felépítéséhez az alapértelmezett, a master és slave jellemzőkkel nem rendelkező helyi mysql szerverek konfigurációs file-ját módosítani kell, a kiindulási állapot adatbázisát a slave mysql szervereket futtató központi physhun gépre el kell juttatni, ott a kiindulási állapot adatait a megfelelő slave mysql szerverbe be kell tölteni, és végül az egyes slave szervereket el kell indítani.

Erre a célra a master-ré váló helyi szervereken futtatni kell a replikaciót a master szervereken beállító *replsetup.pl* script-et, paraméterként a replikálandó adatbázis(ok) nevét kell átadni környezeti vátozóban:

MYSQL_DB='db1 db2 ...dbn' replsetup.pl

A központi physhun gépen a slave mysql szerverekhez az extra konfigurációs file-okat

létrehozandó, és a kiindulási adatbázisokat betöltendő, szükséges a *fromtarsetup.pl* (normál binlog alapú replikáció), illetve *fromdumpsetup.pl* (mysqldump alapú replikáció) script futtatása.
A *fromdumpsetup.pl* és a *fromtarsetup.pl* script-eknek két paramétere van, az első a master mysql szervert futtató gép neve, a második a slave mysql szerver sorszáma, például:

fromtarsetup.pl w3.atomki.hu 6

A projekthez csatlakozók részére a szükséges scripteket rendelkezésre bocsátjuk.

Csatlakozás a PhysHun projekthez

Amint azt a bevezetőben is említettük, a PhysHun egy IHM pályázat keretében készült. A pályázati cél megvalósítása érdekében konzorciumot hoztunk létre. A konzorcium tagjainak (akadémiai kutatóintézetek, egyetemi fizika tanszékcsoportok) képviselõi rendkívül idõszerűnek látják a lokálisan felhalmozódott internetes anyag szélesebb körben való jól struktúráltan áttekinthetővé tételét és az elosztott adatbázis révén megvalósuló biztonságos keresést.

A résztvevõk olyan innovatív technológiát alkalmaznak, amely know-how-ként szolgálhat a továbbiakban a rendszerhez csatlakozni kívánó intézmények számára.

Ezúton - most már a megvalósult pályázaton kívül - meghirdetjük a csatlakozás lehetőségét, s várjuk a fizika tárgyú web-tartalommal rendelkező rendszerek gazdáinak jelentkezését.

A csatlakozáshoz szükséges információk a PhysHun projekt honlapján találhatók.
URL: http://www.kfki.hu/physhun