CINEGE – BIBLIOGRÁFIAI ÉS PÉLDÁNYREKORDOK SZŰRÉSE

Nagy Elemér Károly, eknagy@omikk.bme.hu
Liszkay Béla, bliszkay@omikk.bme.hu
Budapesti Műszaki és Gazdaságtudományi Egyetem
Országos Műszaki Információs Központ és Könyvtár

Egy könyvtár életében a bibliográfiai és példányrekordok formátuma számos alkalommal változik, akarva vagy akaratlanul. Évtizedekig megőrzött, változatlan formátum esetén is előfordulnak formai és tartalmi hibák, nem is beszélve a konverziós hibákról, amelyek a kereshetőséget és a karbantarthatóságot jelentősen megnehezíthetik. Az előadás célja a BME OMIKK-ban és a Universitätsbibliothek der Freie Universität Berlin-ben is alkalmazott, a BME OMIKK-ban fejlesztett szabad Cinege azon részeinek bemutatása, amelyek segítségével a bibliográfiai és a példányrekordok megadott formátumtól való eltéréseit szűrhetjük le parancssoros eszközökkel.

Bevezetés

Azokban a könyvtárakban, ahol hosszú időn keresztül saját építésű adatbázisokkal dolgoznak, az adatbázisok formátuma gyakran változik, egyrészt az adatbázisok technikai hátterét biztosító technológiák és szoftverek változásai miatt, másrészt a adatformátumot érintő szerverezeti, személyi, törvényi és szabványi változások miatt. Mindkét típusú változás következménye lehet az, hogy a régi adatokat konvertálni vagy újrafeldolgozni kell, az adatok mennyisége és a rendelkezésre álló erőforrások miatt azonban tipikusan csak a konverzió a járható út.

A konverziónak tipikusan két problémája van, az egyik a hibás kimenő rekordok problémája, a másik a hibás bemenő rekordok problémája. Amennyiben a konverziós programunk nincs felkészítve explicit a hibás rekordok kezelésére, legjobb esetben is azzal számolhatunk, hogy a hibás bemenő rekordokból hibás kimenő rekordok lesznek. Rosszabb esetben a konverziós program hibaüzenettel leáll (ami egy félórás verziófrissítésből többhetes munkát csinál), elrontja a jó adatok (ha nem vesszük észre időben, akkor visszavonhatatlanul), vagy letörli az egész adatbázist (jó lenne, ha elrettentő példa lenne).

Általánosságban elmondható, hogy minél rugalmasabb egy rekordformátum, azaz minél több értéket illetve értékkombinációt enged meg, annál több hiba kerül előbb vagy utóbb az adatbázisba. A hibák egy részére tipikusan léteznek beépített szűrők, amely az adott hibával rendelkező rekordokat nem engedik elmenteni (vagy legalább figyelmeztetnek), de nem várható el, hogy minden létező kombinációra létezzen ilyen figyelmeztetés, ráadásul a konvertáló programok tipikusan más felületen keresztül érintkeznek az adatbázissal, és ezek a szűrő csak kivételesen működnek együtt a konverziós programokkal. Tipikus esetben tehát marad a konverzió utáni, illetve utólagos szűrés. A bemutatott Cinege egy ilyen utólagos szűrősekre alkalmas eszköz.

Adatszűrési megközelítések

Az adatszűrési megközelítések közül itt és most csak három szempont szerint osztályozzuk az adatszűrőket, mégpedig adatforrás, hiba-reakció és optimizmus szerint.

Az adatszűrők adatforrás szerint például négy csoportra oszthatóak, amennyiben az XML-t külön kategóriaként kezeljük, és nem a szövegfájlok közé soroljuk:

- Szövegfájlból dolgozók
- Adatbázisból dolgozók
- Bináris fájlból dolgozók
- XML fájlból dolgozók

Az adatszűrőket hiba-reakció szerint például négy csoportba oszthatjuk:

- Hiba esetén leálló
- Hiba esetén automatikus javítást megkísérlő
- Hiba esetén hibás rekord azonosítóját feljegyző
- Hiba esetén hibás rekordot kihagyó

Az adatszűrőket optimizmus szerint például két csoportba oszthatjuk:

- A rekordokat alapvetően jónak tekintő
- A rekordokat alapvetően hibásnak tekintő

Természetesen egy-egy adatszűrő a különböző hibákra az említett osztályozások szerint különbözően reagálhat, sőt, általában a használhatóságot javítja, ha beállítható az adatszűrő viselkedése az adott hibára. Így például MOKKA feltöltés során az hírhedt "kalapos õ" automatikusan "ő"-re cserélendő, de a friss rekordok belső ellenőrzése során a hibás rekordok azonosítója későbbi javításra feljegyzendő, lehetőleg a rekordot utoljára megérintő katalogizáló azonosítójával együtt.

A Cinege bibliográfiai és példány szűrők osztályozása

A Cinege a bibliográfiai- és példányrekordszűrői adatbázisból dolgoznak, a rekordokat alapvetően hibásnak tekintik, és hibás rekord azonosítóját feljegyzően működnek. Az adatokat más formátumból letöltő programrészek ezen felül még hiba esetén hibás rekordokat kihagyóan is működnek. Mindkét említett szűrő rugalmasan, parancssorban megadott feltételek alapján szűr, és sokkal nagyobb támogatást ad a hibás rekordokat meghatározó szabályok megfogalmazására, mint a jó rekordokat meghatározó szabályok megfogalmazására, emiatt alapvetően a hibásnak nem talált rekordokat tekintjük jó rekordoknak.

A Cinege bibliográfiai és példány szűrők adatforrásai

Az említett szűrők tipikusan MySQL[1] adatbázisból, JDBC-n keresztül, egyesével szedik a rekordokat, de elvileg bármilyen, JDBC támogatással rendelkező adatbázis használható adatforrásnak. A példány rekordok az *Items* táblából származnak és egy rekord a táblában egy példányrekordnak felel meg, minden mező csomagolatlanul, pakolatlanul és tömörítetlenül szerepel. A bibliográfiai rekordok ezzel szemben a *BibUnits* tábla *BibData* mezejében, csomagolatlanul ás tömörítetlenül, de CSV-be pakolva helyezkednek el, például a HUNMARC-hoz hasonló (mezőkódokkal, almezőkódokkal és indikátorokkal tagolt, ismétléseket megengedő) formátumban.

A reguláris kifejezések

A reguláris kifejezések (regular expressions) körülbelül 1940 óta vannak jelen a műszaki tudományok egy részében és körülbelül 1960 óta az informatikában, a Unixos világban elterjedt *grep* parancs talán a leghíresebb reguláris kifejezést támogató program[2]. Segítségével egyszerűen megfogalmazhatunk például olyan feladatokat, mint a dupla szóközt tartalmazó mezők "`[ ][ ]`", a kötőjelet vagy aláhúzást tartalmazó mezők "`[-_]`" vagy a "Szerk"-el kezdődő és nem szóközre végződő mezők "`^Szerk.*[^ ]$`" vagy például az elmúlt ötvenhat év valamelyikét tartalmazó mezők "`((19[5-9][0-9])|(200[0-5]))`" kiszűrése.

Bibliográfiai és példányrekordok szűrése Cinege-vel

A bibliográfiai és példányrekordok szűrésére az Exporter osztály végzi, a DB2BIB illetve a DB2ADM feladatokkal. A szűrési feltételeket a direktíváknak megfelelően lehet, tipikusan a IF(ANY/NONE/ALL/NOTALL)MATCHES illetve az IFITEMSDATA(MATCHES/NOTMATCHES) alfeladatokkal. Konkrét példákat a 2006-os NetWorkShop előadásban, illetve a méltán világhírű nyílt forrástárházon, a SourceForge-on a https://www.sourceforge.net/projects/cinege címen elérhető jar fájlokban a data/cinege/docs/Examples.pdf fájlokban láthatunk.

1 www.mysql.com
2 Forrás: www.wikipedia.org