XML ALAPÚ BIBLIOGRÁFIAI ADATCSERE A MAGYAR TUDOMÁNYOS MŰVEK TÁRÁBAN

Melléklet az XML export útmutatóhoz

Szövegverzió: 2.1; 2012.11.14.

Tartalomjegyzék

Exportálás XML áganként

Lehetőség van XML-áganként megszabni, hogy mennyi részletet szeretne exportálni. Teljes export esetén minden rendelkezésre álló mező, validációs formátumú export esetén csak az ágra jellemző validációs adatok kerülnek letöltésre, míg "kihagy" esetén az adott ág (alág) kihagyásra kerül. Az "idézők" ágban plusz opcióként lehetőség van „Önálló publikációként” választásra, ekkor minden idéző mint önálló publikáció kerül exportálásra.

Gyakori hibák - rekord javítási lehetőségek

Az export program jól formázott rekordokat állít elő, a hibaellenőrző modul a tartalom validálásánál találhat hibákat. Tartalmi hibák a kézi adatbevitellel és/vagy importtal juthatnak az adatbázisba.

Tartalmi hiba, ha valamelyik mezőben az xml definícióban meg nem engedett karakterek vannak (nem nyomtatható karakterek, valamint a & , < , > , "(idézőjel), ' (aposztróf) karakterek, az xsd séma által meg nem engedettnek definiált mezőtartalom, például érvénytelen (negatív) dátum. Tartalmi hiba az is, ha valamelyik mező nem utf-8-as karaktereket, hanem például a kelet-európai ISO-8859-2-s karaktereket tartalmaz. Lehetnek más problémás karakterek is, amelyek egyes mezőkben okozhatnak nehézséget, a feldolgozó programtól függően.

Módszerek javításra

Az xml hibák javítása informatikai jártasságot igényel. Ha az xml exportmodul hibát talál, akkor a hibás rekordok azonosítóit letölthető fájlba gyűjti és a letöltési lapon a hibás rekordokról ad diagnosztikus információkat, amit a javításhoz jól fel lehet használni.

Hibakereső eljárások

Először a hibásnak minősített rekordot ellenőrzés nélkül érdemes exportálni és elemezni. Az elemzés után lehet törölni vagy javítani az xml definícióknak vagy a mycite.xsd-nek nem megfelelő részeket és a javítás után az xml fájlt újra kell ellenőrizni.

Hibakeresés Notepad++-al

Feltételezve az *XML Tools* plugin telepítését.
Fájl megnyitása -> Plugins (bővítmények) -> XML Tools -> Check XML syntax now - A program a hibás sorra ugrik, ahol rögtön lehet javítani is.

Hibakeresés böngészővel

Firefox (legfrissebb stabil verzió, most 14.0!) Fájl megnyitása Ctrl-O -> ellenőrzés nélkül exportált fájl megnyitása -> hiba helyét a FireFox megmutatja, rámutatva a hibás karakterre, kb. így:
“XML feldolgozási hiba: nem jól formázott. Hely: file:///C:/Users/makaragb/Documents/2_MTMT/_Szoftver%20%C3%BCgyek/Import/xml/106057_ellenorzes-nelkul.xml
1494. sor, 53. oszlop: <reference>Santos, J., (1994) Método Para el An□lisis de la Estabilidad Robusta de Sistemas con Retardo, , Ph. D. Thesis, CINVESTAV-IPN, México, (In Spanish);</reference>”

Chrome (legfrissebb stabil verzió!) Ctrl-O -> ellenőrzés nélkül exportált fájl, a hibára rámutat:
”“This page contains the following errors: error on line 1494 at column 17: Encoding error Below is a rendering of the page up to the first error.”

Javítás az MTMT-ben
Az export letöltő lapján a hibás rekordok megtekinthetők, a hiba lokalizálható rekord szinten. A rekordon belül a lokalizáció néha nehéz. Az MTMT-ben javítást akkor érdemes kezdeni, amikor már a hibá(ka)t a rekordon belül is sikerült lokalizálni.
Az MTMT-be belépve a cikkazonosító segítségével a hibás rekord megkereshető, majd szerkesztéssel a hibás rekord tartalma javítható, hibás dátum javítható, a meg nem engedett karakterek törlendők. Ez a javítási módszer preferálandó, mert a javítás maradandó eredményt ad, későbbi exportoknál nem kell újra javítani. A javításhoz lehet segítséget kérni az MTMT központi adminisztrációjától.

Meg nem engedett karakterek törlése az exportált állományban
Az általános szerkesztő programok (Windows esetében a WordPad, Notepad) ezeket az állományokat szerkeszteni tudják, azonban nem mutatnak rá a hibás rekord hibás részletére. Az xml állomány behívható böngészőbe, ahol a hibák egy részét, a meg nem engedett karaktereket a böngésző azonosíthatja. A Windowshoz letölthető Notepad ++ szerkesztő programban beállítható a rejtett karakterek mutatása is, így az általában “láthatatlan” karakterek is megtalálhatók és törölhetők a hibaellenőrzés nélkül exportált xml állományokban. A Notepad++ az “XML tools” plugin telepítése után jó megoldás a kényelmes hibakeresésre és javításra.
Az állomány javítása után érdemes ellenőrizni, hogy a javított állomány jól formázott és érvényes-e?

Közvetlen export lehetőségek
Az adatbázis adatai a rendszerbe bejelentkezés nélkül, közvetlenül xml (vagy html) formátumban exportálhatók. Az export egyik célja lehet az adatok valamilyen weblapon történő megjelenítése vagy egy külső rendszerben adatfeldolgozás az MTMT-ből származó adatok felhasználásával.

Vázlatos példák további informatikai feldolgozásra:
MTMT ben keresés → találatok listája → xml export → beolvasás xml megjelenítőbe → kiírás docx formátumban → szövegszerkesztés Wordben.
Közvetlen URL-el szerző munkásságának exportálása → mentés → beolvasás xml megjelenítőbe → kiírás docx formátumban → szövegszerkesztés Wordben vagy nyomtatás pdf formátumban
Az MTMT mycite.xsd sémának megfelelő XML fájlok tartalmát megjelenítő szoftver itt található: http//: vm.mtmt.hu/megjelenites

További adatfeldolgozásra alkalmas xml formában a “docres.php” modulon keresztül lehet közvetlen hívással adatexportálást végezni. A docres rendszer használatához a lekérdező számítógépnek az MTMT-nél előre regisztrált ip címmel kell rendelkeznie. A docres rendszer használatához szükséges információkat a regisztrált ip címmel rendelkező felhasználók külön dokumentumban kapják meg.

MTMT - XML tudnivalók

Az MTMT-ből exportált XML állományok tartalma
Az állományok az export beállításoktól függően tartalmaznak a rekordokhoz hozzárendelt szerzőkre, folyóiratokra, forrásközleményekre és idéző közleményekre vonatkozó rekordokat.

Az XML állomány a menteni kívánt lista ÖSSZES, rendezés nélküli idézőjét tartalmazza, tehát az idézőkre vonatkozó lista-beállítások nincsenek hatással az export-állomány tartalmára. Amennyiben válogatott, rendezett idézőket tartalmazó letölthető listára van szüksége, úgy válassza a HTML formátumot.

Az állományok a rekordokat az *azonosítók növekvő sorrendjében* tartalmazzák, mind közleményeket tekintve, mind közleményenként az idézőket tekintve.

A rendszeresen frissített összes böngésző típus rendelkezik automatikus XML formázási ellenőrzéssel, és nem-jól-formázott dokumentum esetén kiírja a hiba tényét és helyét is. Figyelem, ezek a böngészők a mycite.xsd sémaállományt nem használják, a séma definíciók megsértését nem tudják megmutatni!

Amennyiben úgy döntünk, hogy szeretnénk az XML dokumentumokat nemcsak formázási szempontból, hanem validációs szempontból is tesztelni a mycite.xsd állomány felhasználásával, úgy a Microsoft XML Notepad 2007 nevű ingyenesen letölthető és használható programjával elvégezhető a validálás.

Általános XML tudnivalók

A *jól formázott* XML dokumentum

Az XML dokumentum alapértelmezésben Unicode szöveget tartalmazhat, UTF-8 ill. UTF-16 kódolást támogatva. Nem minden UTF-8-as karakter érvényes az xml értelmezésében. Egy helyesen formázott XML dokumentumnak a következő főbb szabályoknak is meg kell felelnie:

- a dokumentumban csak egyetlen egy gyökérelem szerepelhet, amit csak megjegyzések, deklarációk és utasítások előzhetnek meg (pl. a karakterkódolásra vonatkozó információk),
- csak az üres elem (tag) lehet önlezáró, minden más esetben kötelező a nyitó és záró tag használata (természetesen üres elem is lehet nyitó-záró tag-gel ellátott),
- az egyes tag-ek egymásba ágyazhatóak tetszőleges mélységi szintig, azonban ezek nem lehetnek egymást átfedőek, azaz minden nem gyökér-elemet egy másiknak kell magában foglalnia,
- minden tagbeli attribútum kötelezően ’ (aposztróf-) vagy ” (idéző-) jellel kezdődik és ugyan ilyen jellel záródik,
- az XML alapértelmezésképpen Unicode karakterkészletet alkalmaz, melyet opcionálisan definiálni lehet az XML deklarációban vagy esetleg a szállító protokollban, de az ettől való eltérést kötelező definiálni,
- az összes XML részre igaz, hogy case sensitive, tehát megkülönbözteti a kis és nagybetűket (karaktereket).

Érvényes XML dokumentumok

Az XML dokumentum validációjának szükséges, de nem elégséges feltétele, hogy helyesen formázott legyen. A dokumentum helyes formázásán felül kötelezően meg kell felelnie egy típus definíciónak is.

Az MTMT az XML Schemát (XML Schema Definition, XSD) használja adatstruktúra definiálásra. Az XML alapú schema segítségével pontosan és jól deklarálhatóak a felhasználni kívánt adatstruktúrák. Az XSD hátránya az XML alapúságában rejlik: a deklarációt a felhasználó gyakran nehezen vagy egyáltalán tudja olvasva értelmezni.

Az MTMT által használt XML struktúrákat a letölthető mycite.xsd file tartalmazza. Az xsd verziójával egyeznie kell az exportált XML állományban megadott xsd verzió.

Az XML-nek számos olyan tulajdonsága van, amelyek alkalmassá teszik adattovábbításra:

- Unicode karakterkészlet támogatása (így sok speciális karakter gond nélkül megjeleníthető és továbbítható vele),
- mind az ember, mind a gép számára olvasható formátum,
- összetett típusok ábrázolására és kezelésére is alkalmas (fák, listák),
- önleíró formátum, amely struktúra- és mezőneveket ír le speciális értékekkel együtt,
- szigorú szintaktikus és elemzési követelményeket támaszt, ami biztosítja, hogy a szükséges elemzési algoritmus egyszerű, hatékony és ellentmondás-mentes maradjon,
- platformfüggetlen, így bármilyen szoftver környezetben használható,
- letisztult és jól átgondolt szerkezet, mely Internetes szabványokon alapul.

- ingyenes és teljesen szabadon használható,
- sok fejlesztő eszköz áll rendelkezésre, melyek között sok az ingyenesen hozzáférhető,
- hierarchikus struktúrája megfelel a legtöbb dokumentumtípusnak.

Bizonyos alkalmazások szempontjából a következő hátrányokkal rendelkezik:

- bőbeszédű és részben redundáns a szintaxisa,
- (nagyon) összetett dokumentumok olvashatósága már nehézkes,
- nagy tömegű adatátvitel esetén számottevő adatforgalomra kell számítani,
- nincs lehetőség a dokumentum egyes részeinek közvetlen elérésére és frissítésére, erre csak a teljes fa bejárásával van lehetőség,
- az XML relációs és objektum orientált paradigmához kötése néha sok erőfeszítést kíván a programozóktól.

Az XML formátum előnyei és hátrányai

Az adatcserére leggyakrabban használt megoldások (mint pl. RIS, BibTeX, WoS, EndNote, mycite) legfőbb problémája az, hogy egyik sem teljes, nem is elég gazdag, nem bővíthető szabadon mezőkkel. Az univerzális, xml alapú formátum bővíthető, támogatja az adatmozgatást, a komplex adatstruktúrák átvételét (pl. publikációk és idézőik átvételét, elkerülve az idézők téves publikációhoz rendelését), az adatok értékeléseit (MTA Doktori Tanács, OTKA, ODT, doktori iskolák, egyetemek/főiskolák és karaik valamint tanszékeik értékelése, stb.).

A különböző adatbázisok közötti információcserét, illetve külső program(ok) és az adatbázis közötti adatátvételt is lehet xml adatátvitellel megoldani..

Az XML különböző, az alkalmazó által definiálható adattípusok és értékek leírására alkalmas. Elsősorban Interneten keresztüli információ megosztására és tárolására alkalmas formátum, amely jól strukturált, hardver- és szoftverfüggetlen.

Fontos, hogy egy XML dokumentum akkor és csak akkor érvényes (helyes) ha eleget tesz a következő követelményeknek:

- *Jól formázottság* (well formed): egy helyesen formázott XML dokumentum megfelel minden XML szintaxis-szabálynak,
- *Érvényesség* (valid document): egy érvényes dokumentum olyan adatokat tárol, ami megfelel az általunk definiált tartalmi megadási szabálynak (ezeket a mycite.xsd állomány tartalmazza).

Az a dokumentum, ami nem helyesen formázott, nem tekinthető feldolgozható XML-nek, azaz az elemzőnek meg kell tagadnia a feldolgozást.

Exportálás XML áganként

Lehetőség van XML-áganként megszabni, hogy mennyi részletet szeretne exportálni. Teljes export esetén minden rendelkezésre álló mező, validációs formátum esetén csak az ágra jellemző validációs adatok kerülnek letöltésre, míg "kihagy" esetén az adott ág (alág) kihagyásra kerül. Az "idézők" ágban plusz opcióként lehetőség van „Önálló publikációként” választásra, ekkor minden idéző mint önálló publikáció kerül exportálásra.