Metaadatsémák nyilvántartása szemantikus web alapon

Fülöp Csaba, Kovács László, Micsik András
{csaba.fulop, laszlo.kovacs, micsik}@sztaki.hu
Elosztott Rendszerek Osztály
MTA SZTAKI
Magyar Tudományos Akadémia
Számítástechnikai és Automatizálási Kutató Intézet

Kivonat:

A szemantikus web elgondolás alapfeltétele a hálózaton elérhető adatok és metaadatok szemantikai összekapcsolhatósága. Ennek az egyik legkézenfekvőbb módja a digitális könyvtárakban, archívumokban és adatbázisokban használt metaadatsémák összehangolása. Az ezzel kapcsolatos egységesítési, szabványosítási törekvések egyre ígéretesebbek. Kialakulóban van egy RDF alapú séma leíró rendszer, amely alkalmas a metaadatsémák pontos definiálására, és lehetővé teszi a már meglévő sémaelemekből történő építkezést, a sémák újrafelhasználását.

Az MTA SZTAKI Elosztott Rendszerek Osztálya figyelemmel követi ennek a területnek a fejlődését, és szeretné a külföldi eredményeket Magyarországon is hasznosítani, terjeszteni. Terveink között szerepel egy nyílt Internetes szolgáltatás megvalósítása, amely nyilvántartásba veszi az országban használt metaadatsémákat, és lehetővé teszi azok különböző szempontú böngészését, a sémák közti kapcsolatok áttekintését. A szolgáltatás segítséget nyújtana az új sémák összeállításához is, amelyhez a mások által már definiált sémaelemek is felhasználhatók. Egy ilyen szolgáltatás megteremti a lehetőséget a metaadatsémák áttekintésére és újrafelhasználására, amely által a használatban lévő sémák egységesebbé, kezelhetőbbé válnak, illetve elkerülhető lesz mások munkájának megismétlése. A metaadatsémák megosztása és összehangolása országos érdek, hiszen ezáltal válhat gazdaságosabbá az adatok kezelése és hatékonyabbá az adathozzáférés.

Bevezetés

Az MTA SZTAKI Elosztott Rendszerek Osztálya részt vett a CORES[1] európai projektben, melynek célja, hogy segítséget nyújtson a metaadat szerkesztőknek a metaadat szabványok és megoldások együttműködésének növelésével és a különféle metaadat sémák nyilvántartásával. A projekt során elkészítettünk egy metaadat nyilvántartás és egy metaadat sémák létrehozására és újrahasznosítására szolgáló kliens alkalmazást amely a metaadat sémák újrahasznosításához a nyilvántartásban már szereplő adatokat használja. Mind a nyilvántartás és a kliens alkalmazás Szemantikus Web technológiákra épül.

A projekt két építőköve a metaadatok szabványosítása, szabványos leírása valamint a metaadatok internetes erőforrásokhoz történő **egységes** kapcsolása. Az interneten ezek a technikák együtt sokkal hatékonyabbak lehetnek: ma is sok helyen használnak metaadatokat, de nem egységes formában, így az egységes kezelésük sem megoldható

(pl. HTML oldalakban META tagek, mp3 fájlokban id3tagek, word fájlokban binárisan kódolt metaadatok, stb.) valamint hiába adnánk meg egységesen őket, ha a jelentésük, szemantikájuk nem lenne szabványosítva.

Dublin Core

A Dublin Core Metadata Initiative[2] célkitűzése a szemantikus együttműködés támogatása. Hosszú évek óta foglalkoznak a metaadatok szabványosításával és a metaadat sémák megadásának módjával. Az általuk kidolgozott Dublin Core Metadata Element Set[3] egy szándékosan korlátozott metaadat váz amely a 15 legalapvetőbb elemet definiálja. Éppen ennek köszönhetően általánosan használható és számos tudományág körében hasznosnak bizonyult. A Dublin Core és más metaadat-sémák használata folyamatosan terjed, és már jelen van többek között az oktatásban, a könyvtárakban, a közigazgatásban és a szórakoztatóiparban. A tömeges elterjedéssel azonban újabb problémák megjelenése várható.

A Dublin Core (DC) ajánlás elterjedése során kiderült, hogy az alkalmazók azért, hogy területük igényeinek megfelelhessenek, új értelmezéseket adnak az egyes mezőknek, mezőket adnak hozzá és vesznek el. Az ilyen módon átalakított metaadat-sémákat alkalmazási profiloknak[4] nevezték el, és egy szabályrendszert alakítottak ki az alkalmazási profilok elkészítésére.

Szemantikus Web[5]

A Tim Berners-Lee által elképzelt Szemantikus Web célkitűzése, hogy olyan módszert biztosítson, amellyel az internetes erőforrásokhoz egységes módon rendelhetők metaadatok, valamint hogy ezeken a metaadatokon logikai következtetéseket lehessen végezni. Ennek megvalósítására dolgozta ki a W3C a Resource Description Framework[6] (RDF) és az OWL Web Ontology Language[7] ajánlásait.

Az egyes alkalmazásokban használt metaadat sémák leírására szolgál az RDF Schema[8] (RDFS) amely leírja, hogy a metaadat leírás milyen osztályokból építkezhet, adott osztály mely más osztályok kiterjesztése, valamint milyen kapcsolatban állhatnak ezen osztályok példányai, illetve hogy ezek a kapcsolatok hogyan viszonyulnak egymáshoz.

Itt azonban ismét felvetődik az előző részben említett probléma: milyen metaadatokat is használjunk, mik az osztályok és a kapcsolatok? Ilyen szempontból az RDF/RDFS nem mond semmit, ezek csak egy keretrendszert biztosítanak, felhasználásuk módját a felhasználóra bízzák. Ugyan a Dublin Core-nak létezik RDFS leírása[9], de a fent leírtak miatt mindenképpen foglalkozni kell az egyes változatok nyilvántartásával.

Metaadatok nyilvántartása

A különböző metaadat sémákat a szabványosítási szervezetek készítik bizonyos alkalmazási kör igényei szerint. Leírásuk megtalálható az ajánlásaikban vagy az úgynevezett metaadat nyilvántartásokban (metadata registry). Ilyen például a DCMI által üzemeltetett is[10]. Ezen nyilvántartások tartalmát a szabványosítási szervezetek kezelik, a

felhasználók pedig könnyen böngészhetik a sémák elemeit és könnyen felderíthetik az azok közötti összefüggéseket, hierarchiát. Hátrányuk, hogy csak az adott szervezet saját sémáit tartalmazzák.

Azonban sokszor adódhat, hogy a meglévő metaadat sémák nem megfelelőek egy adott alkalmazáshoz annak különleges igényei miatt vagy egyserűen az adott területen még nem történt metaadat szabványosítás. Ekkor rosszabb esetben a fejlesztők saját sémákat készítenek ezzel teljesen inkompatibilissé téve alkalmazásukat más metaadat sémákkal vagy jobb esetben már meglévő sémák részeit használják fel. Ezt felismerve a projektben résztvevő angol partnerünk (UKOLN) már korábbi projektjeiben is egy módszert körvonalazott amelynek segítségével nagymértékben megkönnyíthető az alkalmazás specifikus profilok (application profile) elkészítése már meglévő sémák alapján, illetve a már meglévő profilok újrafelhasználása. A CORES projekt ezen korábbi projektek munkájára építve próbálja meg tökéletesíteni és népszerűsíteni ezt a megoldást.

A módszer lényege, hogy az alkalmazás specifikus profilokat már létező, széles körben elterjedt és elfogadott sémák elemeinek felhasználásával, azok jelentésének megtartásával, tudják elkészíteni az alkalmazás fejlesztők és csak a minimálisan szükséges új elem definiálására legyen szükségük. Így ha alkalmazásuk egy már ismert séma elemeire építik, akkor azt a sémát ismerő többi alkalmazás képes lesz ezen elemek értelmezésére, feldolgozására, míg a többi, újonnan definiált vagy máshonnan származó, számukra ismeretlen elemet figyelmen kívül hagyhatják.

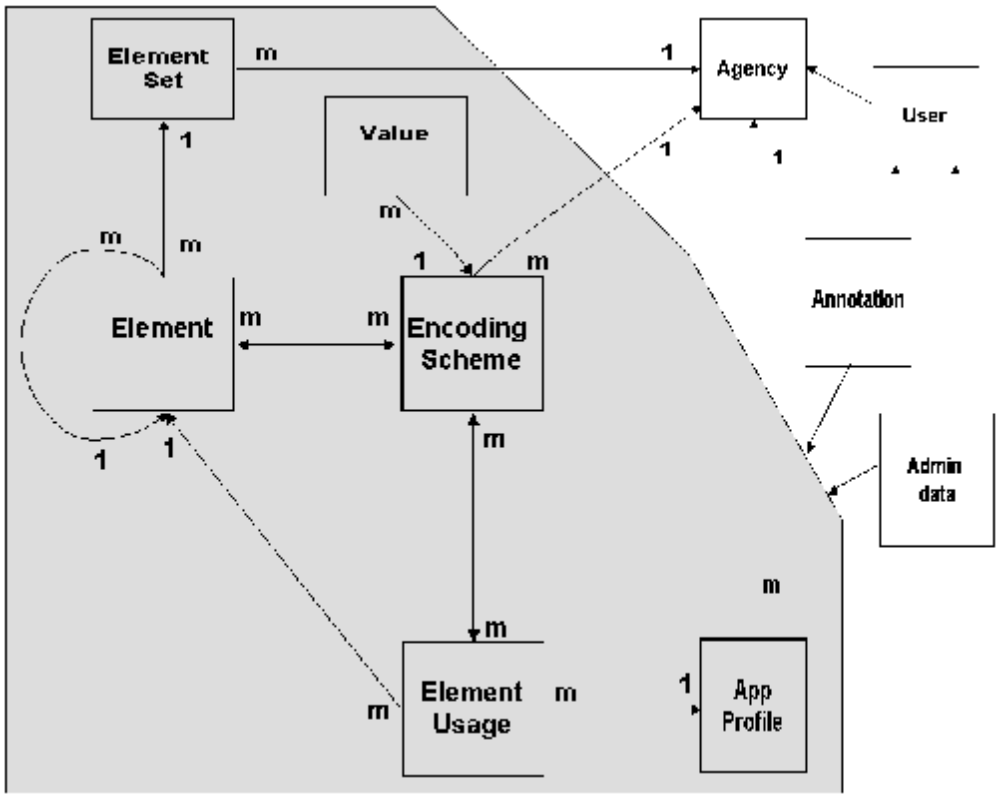
Hogy mindez működhessen, szükség van egy olyan nyilvántartásra, amelyben minél több ismert és elterjedt szabványos séma megtalálható, hogy a profilok készítői kedvükre válogathassanak. Szükség van egy olyan séma és profil szerkesztő eszközre, amely ezen nyilvántartásban képes keresni és lehetőséget biztosít az ott talált elemek újrafelhasználására. A profil elkészülte után pedig lehetőséget az a szerkesztőnek, hogy a saját profilját is feltöltse. Ez a mechanizmus biztosítja hogy mások is rátalálhassanak az adott profilra és azt ők is felhasználhassák: ezzel szorosabb együttműködést biztosíthatnak az alkalmazásaik között, maguknak pedig munkát takaríthatnak meg, hogy nem kell egy új profilt átgondolni és megcsinálni, hanem egy esetleg már jól bejáratottat használhatnak.

A rendszer adatmodellje

A rendszer[11] a következő RDFS modellt adja a metaadat sémák és alkalmazás profilok megadására:

A szürke részben található entitások a metaadat sémák leírásáért felelősek. Az ElementSet és Element entitások a hagyományos metaadat sémák leírására szolgálnak az Encoding Scheme és Value entitások pedig az azokhoz rendelhető megszorítások, minősítők leírására. Az újdonságot az Application Profile és Element Usage entitások jelentik, amelyek az alkalmazás profilok leírására szolgálnak. A profil elemeit az Element Usage-ok adják, amelyek egyértelműen kapcsolódnak egy már létező séma egy eleméhez. Hogy ne csak már létező sémák elemeit lehessen felhasználni ezért az alkalmazások

készítőinek lehetőségük van új sémák létrehozására is, amelyben eddig nem létező elemeket is leírhatnak, majd felhasználhatnak a profiljukban.



A szürke részből kilógó entitások a felhasználói azonosításért és a jogosultságok kezeléséért felelősek, ezzel biztosítva az alapjait egy olyan metaadat nyilvántartásnak amelyet széles közönség használhat. A szervezetek sémáinak és profiljainak a karbantartását az adott szervezethez tartozó felhasználók végezhetik.

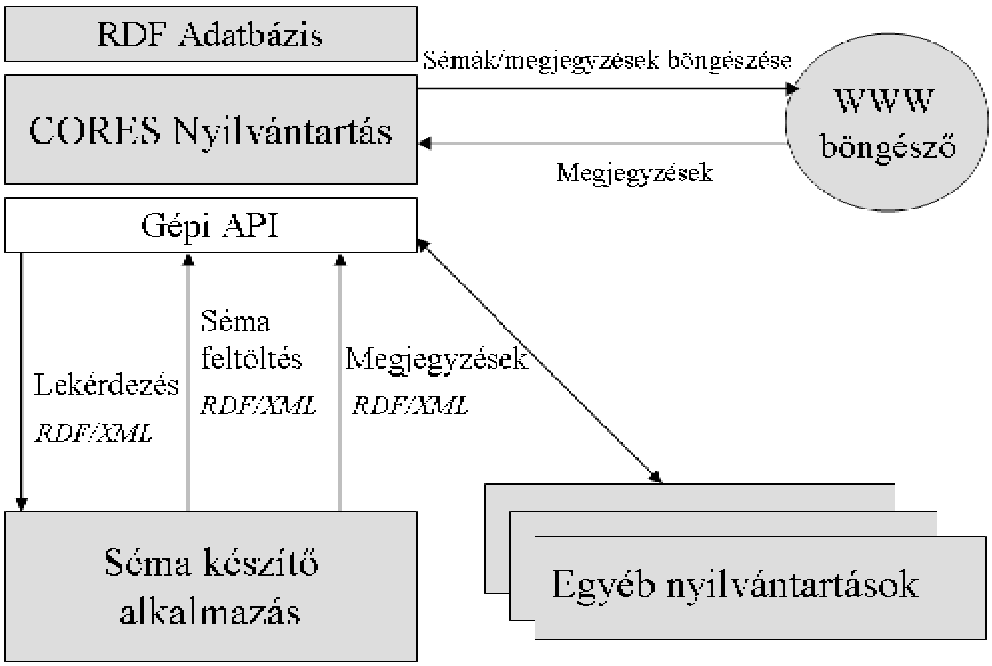
A másik újítása az egyes entitásokhoz fűzhető megjegyzések, kérdések, amelyekben a felhasználók leírhatják tapasztalataikat, ötleteiket az adott entitásról és az azt használó közösség megvitathatja a felmerülő problémákat. A megjegyzéseken túl az egyes sémák fejlődése az adminisztratív metaadatok segítségével is nyomon követhetők, amelyekből kiderül, hogy ki és mikor módosított a sémán vagy profilon[12].

A rendszer architektúrája

A metaadat nyilvántartás egy RDF adatbázisra épülő, perlben írt webes alkalmazás. Ennek feladata az adatbázisában szereplő adatok és azok közötti kapcsolatok megjelenítése, lehetővé téve az azok könnyű felderítését valamint az adatokon végezhető kereső funkció megvalósítása. Eléréséhez csupán egy web böngésző szükséges. Másik feladata egy programok számára elérhető API biztosítása amelyen szintén lekérdezhető és kereshető a tartalma és ezen kívül itt tölthetők fel az új sémák/profilok.

Az új sémák/profilok előállítására egy javaban írt, platform független kliens program használható, amely képes a nyilvántartás API-jával kommunikálni ezzel lehetővé téve, hogy az új profilokat egy folyamatosan bővülő adatbázis alapján készíthessük el.

A rendszer felépítése a következő ábrán látható:



Összefoglalás

A mai weben lévő információk leginkább csak emberek számára értelmezhető formában található meg. A Szemantikus Web a W3C legújabb elképzelése az internet továbbfejlődéséről amely reményeik szerint lehetőséget biztosít a gépek számára is a weben található tartalom értelmezésére és információ kinyerésére. Erre egy jó példa lehet a keresők továbbfejlődése amelyek Szemantikus Web környezetben olyan dolgokat is megtalálhatnak amire a mai full-text keresők nem képesek valamint a sok téves találatot is kizárhatjuk.

Ennek megvalósításához új nyelveket, ajánlásokat dolgoztak ki, amelyekkel metaadatok rendelhetők az internetes erőforrásokhoz és ezeken az adatokon logikai következtetések végezhetők. De ez csak akkor lehet hatékony ha mindenki egységes, szabványosított metaadat sémákat használ az internetes erőforrások leírására. A CORES metaadat nyilvántartása célja ezen egységesítés segítése a már létező sémák nyilvántartásával és azok elemeinek alkalmazás profilokban való újrafelhasználhatóságának biztosításával.

A CORES metaadat nyilvántartás bárki által kipróbálható a http://cores.dsd.sztaki.hu címen, ahol böngészhető a jelenlegi nyilvántartás, valamint rendelkezésre áll egy teszt nyilvántartás is, melyben a séma szerkesztés és feltöltés is kipróbálható.

Köszönet

A szerzők ezúton köszönik a CORES projekt tagjainak segítségét (PricewaterhouseCoopers, Fraunhofer-Gesellschaft , UKOLN, MTA SZTAKI Elosztott Rendszerek Osztály). A CORES projektet az EU támogatja.

---

[1] CORES projekt honlapja: http://www.cores-eu.net/

[2] Dublin Core Metadata Initiative: http://www.dublincore.org/

[3] DC 1.1: http://dublincore.org/documents/dces/

[4] Rachel Heery, Manjula Patel: *Application profiles: mixing and matching metadata schemas,* http://www.ariadne.ac.uk/issue25/app-profiles/

[5] Szemantikus Web: http://www.w3c.org/2001/sw/

[6] RDF: http://www.w3.org/RDF/

[7] OWL: http://www.w3.org/TR/owl-ref/

[8] RDFS: http://www.w3.org/TR/rdf-schema/

[9] DCMI Schemas: http://www.dublincore.org/schemas/

[10] DCMI Registry: http://dublincore.org/dcregistry/

[11] CORES metaadat nyilvántartás: http://cores.dsd.sztaki.hu

[12] Rachel Heery, Pete Johnston, Csaba Fülöp, András Micsik: *Metadata schema registries in the partially Semantic Web: the CORES experience,* http://www.siderean.com/dc2003/102_Paper29.pdf