Rendező:

Hungarnet

Networkshop 2020 ONLINE

Házigazda:

PÉCSI TUDOMÁNYEGYETEM
UNIVERSITY OF PÉCS

„Esélyeink és kihívásaink a digitális transzformáció világában”

Országos Online Konferencia
2020. szeptember 2–4.

INNOVÁCIÓS ÉS TECHNOLÓGIAI MINISZTÉRIUM

Szerkesztette: Tick József, Kokas Károly, Holl András

Tipográfia és tördelés: Vas Viktória

Networkshop
2020. szeptember 2-4. Pécsi Tudományegyetem, (On-line)
konferencia előadásainak közleményei

ISBN 978-615-01-0376-1
DOI: 10.31915/NWS.2020

Kiadja a HUNGARNET Egyesület
az MTA Könyvtár és Információs Központ közreműködésével
Budapest
2020

Borítókép: freepik.com

DOI: 10.31915/NWS.2020.8

A mikroadatok felhasználása webarchiválási környezetben

Németh Márton
web könyvtáros
Országos Széchényi Könyvtár, Webarchiválási Osztály

doktorjelölt
Debreceni Egyetem, Informatikai Tanulmányok Doktori Iskola

Ebben a rövid tanulmányban először rövid áttekintést szeretnék nyújtani a mikroadatokkal ellátott szemantikus jelölőkről s azok általános könyvtári felhasználásáról. A továbbiakban néhány a webarchiválást érintő kihívást említek meg az archivált tartalom megjelenítése, a kutatástámogatás és a hosszútávú digitális megőrzés témaköreihez kapcsolódva. A tanulmány fő célja annak vizsgálata, hogy a mikroadatok felhasználása miként tud segíteni e kihívások kezelésében. Fontosnak tartom kiemelni az International Internet Preservation Consortium (IIPC) Research Working Group[1] tevékenységét, mely nemzetközi értelemben is segíthet összehangolni az érintett kutatási és gyakorlati felhasználási tevékenységeket.

1. Mikroadatok a dokumentumalapú web és a szemantikus web metszéspontjában

A HTML szabvány megszületése óta néhány nagyon alapvető metaadat elem (pl. a title – cím elem) adható hozzá egy adott honlap forráskódjának fejlécéhez.[2] Az adott weboldalhoz csatlakozó robots.txt nevű fájlban adhatjuk meg, hogy a webaratást végző robot milyen tartalmakat menthet le, s a honlap mely részeiből van kitiltva. Ezek a szolgáltatások a dokumentumalapú webhez kötődnek, ahol az adott weboldal témájáról csupán igen felszínes információkhoz juthatunk hozzá. Nem kapunk segítséget ahhoz, hogy a honlap tartalmi jellemzőiről pontos képet kapjunk, s ahhoz sem, hogy miként kapcsoljunk össze egymással különböző honlapokat tartalmi jellemzőik szerint. A dokumentumalapú web dokumentumokat köt össze, minősítés nélküli hivatkozásokkal. A szemantikus web bizonyult a következő fejlődési lépcsőfoknak[3], mely nem a különálló dokumentumegységekre (weboldalakra) hanem azok tartalmára fókuszál, hogy miként lehet az egyes tartalmi elemeket meghatározni, leírni majd összekötni egymással a globális információs környezetben. A szemantikus web dokumentumok helyett különféle adatkészleteket köt össze minősített linkek segítségével. Ennek révén a web alapú tartalmi erőforrások egységes platformon válhatnak kezelhetővé a könyvtári katalógusokból, vagy bármely információforrásból származó szabványos módon feldolgozott adatokkal. A mikroadatok[4] hidat képeznek a dokumentumalapú, illetve a szemantikus web között. Lehetővé válik, hogy szemantikus kijelentéseket tegyünk, egy honlap forráskódjába

1 Research Working Group IIPC, „Research Working Group – IIPC", IIPC Research Working Group, 2020, https://netpreserve.org/about-us/working-groups/research-working-group/.
2 „HTML 5.2", HTML 5.2 W3C Recommendation, 14 December 2017, 2017, https://www.w3.org/TR/html52/.
3 „Scientific American: The Semantic Web", elérés 2020. augusztus 14., http://web.archive.org/web/20070713230811/http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.
4 „HTML Microdata", elérés 2020. augusztus 14., https://www.w3.org/TR/2018/WD-microdata-20180426/.

ágyazva, ily módon beillesztve az honlapot a szemantikus web univerzumába. Erről Horváth Ádám tartott a 2016. évi Networkshopon áttekintő előadást[5]. A mikroadatok a webes szabványokhoz és szemantikus webes adatmodellekhez kötődnek tehát. Az RDFa (or Resource Description Framework in Attributes) egy W3C konzorciumi ajánlás mely attribútum szintű kiterjesztéseket nyújt a HTML, XHTML és különféle XML alapú dokumentumtípusokhoz, gazdag metaadatokkal ellátva azok elemeit.[6] Ez az első fontos lépés, mivel az RDFa az RDF adatmodell része, mely a szemantikus web egyik alapvető építőköve. A szemantikus metaadatok beillesztésével alany-állítmány-tárgy típusú kijelentések ágyazódnak be a fent említett típusú dokumentumokba. Ennek legnagyobb előnye, hogy az RDF kijelentések állítmányi részének elemei különféle szemantikus szótárakból is származhatnak. Így különféle sémák, illetve platformok azonos platformon használhatók fel. Az egyik legfontosabb szótár a schema.org. Ezt a Bing, Google és a Yahoo közösen kezdte fejleszteni, hasonlóan néz ki, mint egy általános tezaurusz.[7] A legfontosabb hozzáadott értéke abban áll, hogy lehetővé teszi bármilyen, a weben megjelenő HTML oldal tartalmi leírását.

Miért fontosak a megfelelő szótárakra támaszkodó mikroadatok? Az egyik legfontosabb szempont ennek kapcsán a pontosság. A schema.org jelölőkkel címkézett HTML oldalak sokkal értelmezhetőbbé válnak a keresőmotorok számára, így azok pontosabb találatokat tudnak visszaadni. A weboldalak indexelése is sokkal könnyebbá válik a szemantikus mikroadatok révén, így az azokat használó tartalomforrások fontossága is felértékelődik. A schema.org felhasználásának egyik úttörője s azóta is markáns szereplője a könyvtári integrált rendszereket fejlesztő cégek közül az Online Computer Library Center (OCLC). Minden leírás a WorldCat katalógusrendszerükben schema.org jelölőket is tartalmaz.[8] Ennek eredményeként az onnan származó rekordok a keresőmotorok találati listáinak élén jelennek meg. A kereskedelmi szereplők mellett a nyílt forráskódú integrált könyvtári rendszerek fejlesztésében is követik az OCLC gyakorlatát. A mikroadatok felhasználása egyébként nem csupán az indexelés pontosságának növelése miatt kerül előtérbe, hanem azért, mert a felcímkézett weboldalak a nyílt kapcsolt adatok univerzumának részévé válhatnak. A szemantikus adatkészletek egymással történő összekapcsolása az információ visszakeresést eddig soha nem látott adattömegre támaszkodva teszi nagy pontossággal lehetővé. Emellett azt is könnyen beláthatjuk, hogy a mikroadatok révén szemantikus címkékkel ellátott digitális dokumentumok hosszútávú megőrzésének hatékonysága is nő, azzal a hozzáadott értékkel, melyet a mikroadatok felhasználása nyújt.

E bevezető után a következőkben rátérünk arra a cikkünk tárgyául szolgáló szűkebb, fő témára, hogy a mikroadatok felhasználása milyen előnyöket rejthet a webarchiválás területén.

5 Horváth, Ádám RDFa – schema.org: a dokumentum web és a szemantikus web egyesítése (Networkshop 2016, Debrecen, 2016), https://conference.niif.hu/event/5/session/10/contribution/27/material/slides/0.ppt.

6 „XHTML+RDFa 1.1 - Third Edition", 2015, https://www.w3.org/TR/xhtml-rdfa/.

7 „Getting Started – schema.org", 2020, https://schema.org/docs/gs.html.

8 „WorldCat Linked Data Vocabulary | OCLC Developer Network", elérés 2020. augusztus 14., https://www.oclc.org/developer/develop/linked-data/worldcat-vocabulary.en.html.

2. A mikroadatok felhasználásának lehetőségei a webarchiválás kapcsán

A schema.org könyvtári területre történő kiterjesztését 2015-ben tették közzé.[9] Ez a kiterjesztett elemkészlet együtt az RDFa séma elemekkel rengeteg integrált könyvtári és repozitóriumi rendszerbe került már integrálásra (Koha, VuFind, Islandora, Dspace stb.).

Ennek révén hatalmas tartalommennyiség vált a korábbinál pontosabban feltárhatóvá és arathatóvá, állandó URL címek, megfelelően összeállított webhelytérképek és robots.txt beállítások és persze természetesen a schema.org szabványába foglalt jelölési elemek révén. Korábban arathatatlannak tartott teljesszövegű adatbázisok bukkantak így fel a felszíni webre. Az aratást ráadásul nagymértékben segítik azok a minősített linkek melyek az egyes tartalomelemek (rekordok) közötti bibliográfiai kapcsolatokat tárják fel.

Másfelől azonban sajnos hasonló schema.org kiterjesztés nem áll rendelkezésre egyéb webes információforrások archiválásának segítésére. Az OCLC kibocsátott egy webarchiválási metaadat irányelv készletet[10], mely tartalmazza a felvázolt (főként Dublin Core alapú) adatelemek schema.org megfeleltetését is. Ez azonban csak egy nagyon korlátozott számú és hatókörű elemkészletet foglal magában. A schema.org, illetve bármely egyéb szemantikus szótár nem rendelkezik olyan speciális elemkészlettel, mely a webarchiválás hatékonyságának növelését biztosítaná. Egy új séma megalkotása, illetve egy meglévő szótár elemkészletének kiegészítése további kutatásokat igénylő feladat. Ebben a tanulmányban arra vállalkozunk, hogy felvillantsunk néhány területet, ahol a szemantikus alapú mikroadat jelölőelemek használata számottevő előnyökkel járhatna.

2.1 A webarchiválással kapcsolatos kihívások

Mielőtt rátérnénk a mikroadatok felhasználási lehetőségeire a webarchiválás környezetében, ebben a fejezetben felvázolunk néhány olyan kihívást melyek kapcsán a mikroadatok felhasználása szóba kerülhet.

Az első jelentős kihívás, hogy sok esetben az aratórobotok nem tudják megfelelő módon feltárni az archiválandó weboldalak szerkezetét mert az nincs egyértelműen meghatározva. További problémát jelenthetnek olyan elavulóban lévő tartalomformátumok, amelyeket a napjainkban használt böngészők már nem is támogatnak. Sokszor nem is egyszerű meghatározni, hogy milyen fajta formátum, illetve kiegészítő szoftver használatát írták elő az adott weboldal fejlesztői. Azzal könnyen szembesülhetünk, hogy bizonyos tartalmak hiányoznak egy archivált weboldalról, de az okok nagyon összetettek lehetnek s a számítógépes világháló mint tartalomhordozó eszköz komplexitása tükröződik voltaképen ebben vissza. Egy másik jelentős kihívás a nagymennyiségű adatok kezelése. Az ezekből összeálló szintén jelentős méretű adatkészletek kutathatóvá tétele is összetett feladat. A megfelelő mikroadat elemkészletek meghatározása és használata többek között ezeken a területeken is segítséget nyújthatna. A továbbiakban ehhez szolgálunk még újabb adalékokkal.

9 „Bib.schema.org-1.0 – Schema Bib Extend Community Group", 2020, https://www.w3.org/community/schemabibex/wiki/Bib.schema.org-1.0.

10 Dooley Jackie és Kate Bowers, „Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group", 2018, https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata/recommendations.html.

2.2 A robotokkal végzett webarchiválás segítése mikroadatok révén

Egy leeendő mikroadat jelölőkészlet néhány eleme felhasználható lenne az aratórobot munkájának segítésére egy adott honlap szerkezetének feltárása kapcsán. Az oldaltérkép XML formátumban épülhetne fel. A robots.txt fájlba pedig különleges parancsok lennének beilleszthetők melyek a robotok tevékenységét szabályoznák. A robots.txt fájlba foglalt információk nem csupán a gyökérkönyvtárban található önálló fájlban jelenhetnének meg, hanem az egyes oldalak fejléceibe is be lehetne illeszteni egy részüket, a megfelelő hozzáférés biztosítása végett. A robots.txt fájlban vagy a honlapfejlécben speciális jelölőkkel rögzített információtípusok egy része jelenleg egy nem szabványosított kiegészítő elemkészlet[11] használatával határozható meg. Ezek közé tartozik például a noindex metacímke, noindex http válaszfejléc, illetve további utasítások, mint az allow, sitemap, host vagy az universal match. Ezek jelenleg nem részei a hivatalos szabványnak, azonban rögzíthetők lennének abban, lehetővé téve használatukat az összes weboldalon és minden platformon. Összeállítható lenne ily módon például egy olyan honlaptérkép melyben kizárhatók lennének az aratásból speciális jelölők révén az örökнaptár alkalmazások, adatbázisokra vezető linkek, megelőzve azt, hogy csapdába kerüljön a robot, ahonnan nem is tudna kilépni. Ezek a kivétel korlátozások egy adott oldal fejlécében a noindex http válaszfejléc segítségével lehetnének kezelhetől. A robotok kizárását szabályozó szabvány további kiterjesztései lehetőséget adnának egy honlap különböző felületeinek meghatározására (mobilbarát felület, akadálymentes felület, robotbarát felület, alapértelmezett PC-n megjelenő felület). Meg lehetne határozni, hogy a robotok az összes, vagy csak egy meghatározott felülethez férhetnének hozzá, az archiválási szabályzatban előírtak szerint.

Mikroadatok segítségével a hirdetési felületek, a felugró figyelmeztetések, kötelezően közzéteendő tartalomelemek is felcímkézhetők lehetnének. Az intézményi archiválási szabályzat pedig kitérhetne arra is, hogy ezen elemek közül melyek tartozzanak bele annak hatókörébe.

Nemcsak azt lehetne szabályozni, hogy milyen típusú honlapfelületek lehetnének arathatók, hanem a különféle nyelvi verziókat is el lehetne különíteni egymástól s dönteni, hogy melyiket archiválnánk. Amennyiben egy honlap kezdőoldala egy kereső dobozt tartalmaz csupán, a robot onnan nem tud továbblépni a belső tartalmak felé. Ezért hát speciális jelölők segítségével a robot számára meg lehetne határozni, hogy melyik oldal legyen az aratás kiindulópontja, ahonnan az adatrekordok kinyerhetők lennének. Ily módon, habár a lekereső felület nem lenne a maga funkcionális teljességében archiválható, a mögötte lévő adatok viszont igen!

Az imént felsorolt információk szabványos módon történő megjelölése mikroadat jelölők révén látványosan növelhetné a jelenleg használatos webarchiváló eszközök hatékonyságát, elkerülhetővé téve további költséges szoftverfejlesztési projektek kivitelezését.

11 „Robots exclusion standard – Wikipedia", 2020, https://en.wikipedia.org/wiki/Robots_exclusion_standard.

2.3 Az archivált webes anyag megjelenítések segítése mikroadatok révén

Nagyon sokszor szembesülünk azzal, hogy habár az archiválás sikeres volt, a tárolt anyag nem nézhető vissza annak teljességében a visszanézést biztosító szoftver hiányosságai miatt. Speciális mikroadat jelölők lehetnének a segítségünkre az archivált anyag szerkezetének és felületének értelmezésében is. Például sokszor speciális JavaScript vagy pdf beépülőket használnak a honlapba illesztett fotók, illetve pdf fájlok megjelenítéséhez, illetve lehulló menüből tárul fel az adott honlap szerkezete. Amennyiben speciális mikroadat jelölők révén meg tudnánk határozni ezeknek a fotóknak, fájloknak, menüknek az alternatív megjelenési módját, az archivált tartalom teljes mértékben visszanézhető lehetne.

2.4 Hosszútávú megőrzés támogatása mikroadatok segítségével

A mikroadat jelölők egy csoportja segíthetne meghatározni az adott weboldalon használt tartalomformátumok típusait, verzióit, s egyéb fontos jellemzőit. Rövid kijelentésekkel fel lehetne tárni az összes, az adott oldalon használt, szoftver jellemzőit. A hosszútávú megőrzés szempontjából kihívást jelentő komponensek pl. a beépülő Java vagy Flash kisalkalmazások (appletek) is elkülönítve feltárhatók lehetnének verziószámaik és funkcióik szerint. A jövőben ezek az adatok különleges fontossá válhatnak amikor a tartalom konverzióját, illetve emulálását kell megtervezni egy új archiválási-szolgáltatási rendszerben.

2.5 Kutatástámogatás mikroadatok segítségével

A webarchívumok kutatási célú használatának támogatása mindenhol nagyon fontos összetevője a webarchiválási szolgáltatási környezetnek. A szemantikus adatok szolgáltatása egyrészt a hatékony információ visszakeresést tudja segíteni a kutatók számára. Másrészt a nyílt kapcsolt adatok univerzumának segítségével a különféle adatkészletek között korábban fel nem tárt összefüggésrendszerekre is összpontosíthat a kutatás. A legfontosabb ebben az értelemben azoknak a leíró jellegű attribútumoknak a használata, melyek a hiperlinkekhez társíthatók feltárva azok típusát, illetve a forrás és a cél erőforrás közötti kapcsolat jellemzőit. A szemantikus weben az RDF típusú linkek alapvető fontosságúak a nyílt kapcsolt adatok környezetében a különféle adatkészletek közötti kapcsolat megteremtése kapcsán. A kapcsolat jellemzői az RDF szerkezetben az állítmányi részben fogalmazhatók meg. A feltárt kapcsolati jellemzők pedig automatizáltan gépi úton feldolgozhatók a szemantikus web globális gráfjában.

Az RDF adatszerkezetben az rdf:type tulajdonság adja meg a különféle szótárak, sémák által meghatározott kapcsolódási módokat (kijelentéseket vagy meghatározásokat) a különféle nyílt kapcsolt adat alapú adatkészletek között.[12] Például a weboldalakon használt személynevek vagy földrajzi nevek leszűrhetők és felcímkézhetők lehetnének különféle névtér sémák és azonosítók szabályai szerint. Az adott névalakhoz kötődő a szemantikus Wikipédiára, vagy szemantikus könyvtári katalógusok megfelelő tételeire történő utalás is elérhető lenne ily módon.

12 „link relation types – Microformats Wiki", 2020, http://microformats.org/wiki/link-relation-types.

Egy másik fontos kutatástámogatási célú felhasználási területe a szemantikus jelölők használatának egy adott honlap létrejöttének, illetve módosítási dátumainak rögzítése lehetne. Nagyon fontos az archivált anyag, mint történeti forrás hitelességének biztosítása. Sajnos néhány napjainkban használatos visszanéző szoftver egymásra csúsztat különféle időpontokban archivált tartalmi elemeket a megjelenítés során. Az adott webhelyek, weboldalak, vagy akár egyes weboldalrészek elkészítési, illetve módosítási időadatainak megjelölése mikroadat jelölőcímkék segítségével jelentősen megkönnyítené a történetileg is hiteles összkép felvázolását.

A mikroadat jelölők a digitális bölcsészeti kutatások támogatásában is segítséget nyújthatnának. A digitális filológusok számos XML alapú elemző eszközt és eljárást használnak. Ennek kapcsán szintén szóba kerülhetne a megfelelően címkézett szemantikus szótári (névtér) elemek azonosítása. Például egy magyar szövegben megbúvó latin vagy görög nyelvű idézet, vagy egy adott személy, illetve földrajzi hely különféle névalakjai, különféle helyesírású történeti névalakok is egységesen azonosíthatók lennének ily módon.

2.6 A mikroadatok előállítása és alkalmazása

Amikor a különféle mikroadat jelölőkészletek felhasználási lehetőségeiről esik szó, elkerülhetetlen a kérdés, hogy ki állítja elő ezeket a szabványos adatkészleteket önálló szótár vagy meglévő sémához, névtérhez illesztés formájában? S ki mutat utat a leendő jelölési elvek alkalmazásának terén? Egyértelműnek tűnik, hogy a főszerepet e téren a jelentős archivált webes gyűjteménnyel bíró közintézményeknek, illetve a tartalomszolgáltatói piac főszereplőinek kellene vállalniuk (a legnépszerűbb blogfelületek, közösségi média platformok stb.). Ezek a szereplők megfelelő erőforrásokkal rendelkeznek ahhoz, hogy megalkossák a megfelelő jelölőkészleteket rögzítő szótárakat és előmozdítsák azok felhasználását is. A webarchiválás területén érdekelt intézményeket összefogó nemzetközi konzorcium, az IIPC fontos szerepet tölthetne be abban, hogy a tartalomipar szereplői és a közintézmények egymásra találjanak, s a megfelelő jelölőkészletek illetve szótárak kidolgozása mellett azok használatát a piaci terület szereplői beépítsék a szolgáltatási kereteik közé is. A jelölők használatának elterjedése és a webarchiválás hatékonyságának látványos növelése ilyen módon lenne biztosítható.

2.7 Zárszó

Ez a tanulmány a szemantikus mikroadat alapú jelölők használatának lehetőségeit járta körbe a webarchiválás hatékonyságának növelése kapcsán. A szerző abban reménykedik, hogy a kihívás fontosságát tekintve új kutatás-fejlesztési projektek indulnak majd széles partnerségi keretek között. Ezek eredményeinek a tartalomszolgáltatási gyakorlatba való átültetése garantálhatná igazán a webarchiválás sikerét

Bibliográfia

„Bib.schema.org-1.0 - Schema Bib Extend Community Group”, 2020. https://www.w3.org/community/schemabibex/wiki/Bib.schema.org-1.0.

„Getting Started - schema.org”, 2020. https://schema.org/docs/gs.html.

Horváth, Ádám. „RDFa - schema.org: RDFa - schema.org: a dokumentum web és a szemantikus web egyesítése. Előadás Networkshop 2016, Debrecen, 2016. https://conference.niif.hu/event/5/session/10/contribution/27/material/slides/0.ppt.

HTML 5.2 W3C Recommendation, 14 December 2017. „HTML 5.2”, 2017. https://www.w3.org/TR/html52/.

„HTML Microdata”. Elérés 2020. augusztus 14. https://www.w3.org/TR/2018/WD-microdata-20180426/.

IIPC, Research Working Group. „Research Working Group - IIPC”. IIPC Research Working Group, 2020. https://netpreserve.org/about-us/working-groups/research-working-group/.

Jackie, Dooley, és Kate Bowers. „Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group”, 2018. https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata/recommendations.html.

„link relation types - Microformats Wiki”, 2020. http://microformats.org/wiki/link-relation-types.

„Robots exclusion standard - Wikipedia”, 2020. https://en.wikipedia.org/wiki/Robots_exclusion_standard.

„Scientific American: The Semantic Web”. Elérés 2020. augusztus 14. http://web.archive.org/web/20070713230811/http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.

„WorldCat Linked Data Vocabulary | OCLC Developer Network”. Elérés 2020. augusztus 14. https://www.oclc.org/developer/develop/linked-data/worldcat-vocabulary.en.html.

„XHTML+RDFa 1.1 - Third Edition”, 2015. https://www.w3.org/TR/xhtml-rdfa/.