Online folyóiratcikkek ötcsillagos értékelési keretrendszere

Szemantikus web

Tim Berners-Lee, a web megalkotója, 2009-ben egy ötfokozatú skálát javasolt a szabad hozzáférésű adatok minősítésére. Az összekapcsolt nyílt adatállományok (Linked Open Data) a szemantikus web lényeges elemei, ezért célszerű minél „intelligensebb" formában közzétenni őket. A javaslat szerint csillagok szimbolizálják az egyes szinteket:

- ★ Tedd közzé egy szabad hozzáférést biztosító licenc alatt az adatokat a weben (tetszőleges formátumban).
- ★★ Számítógéppel értelmezhető, strukturált formátumban publikáld őket (pl. egy Excel fájlban, ahelyett, hogy képként beszkenneled a táblázatot).
- ★★★ Ugyanaz, mint a kétcsillagos szint, de válassz nyílt formátumot (pl. Excel helyett CSV-t).
- ★★★★ Ezenkívül használd a W3C nyílt szabványait (RDF és SPARQL) az adatok jelentésének beazonosítására, hogy mások hivatkozni tudjanak az anyagodra.
- ★★★★★ A fentieken túl kapcsold össze az adataid másokéival, hogy kontextusba kerüljenek.

Bár a szemantikus web hívei időnként az ótestamentumi prófétákra emlékeztetnek, akik hiába hirdették az igazságot a népnek, az utóbbi időben már érezhetően egyre szélesebb körben kezdik elfogadni és támogatni ezt a víziót. Köszönhető ez egyrészt az olyan nagy befolyású partnerek megnyerésének, mint amilyen a BBC, másrészt pedig annak, hogy a szemantikus web alapelveit ügyes marketinggel a Linked Data projekt (*linkeddata.org*) zászlaja alatt is terjesztik. Ezek az alapelvek egészen egyszerűek: ha a dolgokat és a köztük levő relációkat számítógéppel értelmezhető módon azonosítani és definiálni tudjuk egyedi URI hivatkozásokkal, melyek nyilvános és általánosan elfogadott, strukturált szótárakra (ontológiákra) mutatnak, valamint ha minden egyes kapcsolat leírható egy egyszerű alany-állítmány-tárgy hármassal *(triple)* az RDF (Resource Description Framework) szintaxist követve, akkor az így megfogalmazott állítások információs hálóvá kapcsolhatók össze (RDF gráf), melyben az eredeti állítások igazságtartalma megőrződik, így tudásháló, vagyis szemantikus web jön létre. Az ontológiákban rögzített leírások lehetővé teszik, hogy egymástól független forrásokból integráljunk adatokat anélkül, hogy elveszítenénk vagy kétértelművé tennénk a jelentésüket. Az egyszerű XML-alapú adatcserének ugyanis megvan az a hátránya, hogy a jelölő címkéknek nincs univerzálisan elfogadott jelentése, így a szinonimák esetén bizonytalanság lép fel (pl. az egyik rendszerben használt *creator* címke vajon azonos-e egy másik sémában a *composer* vagy a *choreographer* címkékkel?), a homonimák pedig félreértésekhez vezethetnek (pl. a *gift* címke jelentése egy angol nyelvű adatbázisban *ajándék*, egy németben viszont *méreg*).

Ma már meggyőző példák léteznek a szemantikus web technológia előnyeire. Az egyik ilyen az ókori művészetet bemutató CLAROS (*clarosnet.org*), ahol eltérő metaadatsémákat használó múzeumi és tudományos adatbázisokból származó információkat sikerült egy egységes rendszerré integrálni.

A könyvtárak és a szemantikus web témájával foglalkozó „Semantic Web in Libraries" nevű éves találkozónak 2011 novemberében Hamburg adott otthont (swib.org/swib11/), melynek fő témája a webes környezetben zajló tudományos kommunikáció volt.

Szemantikus publikálás

A kutatási eredmények nyilvánosságra hozatalának és terjesztésének mindmáig a legfontosabb eszköze a folyóirat. A tudományos folyóiratcikk formája nem sokat változott az elmúlt mintegy 350 év alatt: lényegében egy lineáris szerkezetű beszámoló, melyben a szerző igyekszik meggyőzni olvasóját a hipotézisének helyességéről, melyet

egy nagyobb adathalmazból kiválasztott bizonyítékokkal támaszt alá. Manapság már a folyóiratkiadók többsége PDF formátumban is közzéteszi a publikációkat, ám mivel ezek a fájlok csupán a nyomtatott cikkek digitális hasonmásai, nem tartalmaznak szemantikus vagy interaktív elemeket, alig értelmezhetők a számítógépek számára, és alkalmatlanok arra, hogy automatikus módszerekkel össze lehessen kapcsolni őket más cikkekkel vagy gazdagítani a tartalmukat. Különféle próbálkozások léteznek a tudományos kommunikáció ezen klasszikus formájának megreformálására, a web által nyújtott lehetőségek jobb kihasználására. Egyesek a cikkek HTML verzióinak szemantikus jellegű feljavításával kísérleteznek; mások szövegbányászati szolgáltatásokat fejlesztenek, amelyekkel automatikusan lehet szemantikus jelölőket rendelni a HTML fájlokban található névelemekhez; vagy éppen olyan böngésző-kiegészítőt készítenek, amely képes a hivatkozott publikációkból megjeleníteni részeket anélkül, hogy el kellene hagynunk az eredeti cikket. Van már „okos" PDF-olvasó is (pl. az Utopia Documents), amely annotációs rétegeket ad hozzá az amúgy statikus PDF dokumentumokhoz és így élővé, interaktívvá teszi őket. A szemantikus publikálás és hivatkozás támogatása céljából megalkották a SPAR (Semantic Publishing and Referencing) nevű ontológiát (purl.org/spar/). Egyes kiadók pedig saját modelleken dolgoznak: ilyenek például a *Royal Society of Chemistry* „Project Prospect" vagy az *Elsevier* „Article of the Future" projektjei, illetve a *Pensoft* által kiadott élettudományi elektronikus folyóiratok.

A szemantikus publikálás olyan webes és szemantikus webes technológiák alkalmazását jelenti, amelyek:

- gazdagítják az elektronikusan publikált cikket – például interaktív ábrák, átrendezhető irodalomjegyzékek, szemantikus „nagyítók" (ezek grafikonná alakítanak egy táblázatot, vagy animációvá egy diagramot, amikor rávisszük az egérkurzort);
- gazdagítják a cikk tartalmát – például a szövegben előforduló névelemek szemantikus jelölésével, melyek így linkként szolgálnak a szakkifejezések és fogalmak definíciójához, illetve lehetővé teszik további információk lekérését (pl. a cikkben említett fehérje nevét összekapcsolják a tulajdonságait leíró Protein Database rekorddal);
- elvezetik az olvasót más releváns forrásokhoz – például a szerző honlapjára, vagy vegyszerbeszállítók katalógusaihoz, vagy nemzetközi szervezetek weboldalaira (ilyen lehet mondjuk a WHO honlapjára mutató link egy járványtani témájú cikkben);
- közvetlen linkeket adnak minden hivatkozott publikációra;
- szerkeszthető módon teszik hozzáférhetővé a cikkben szereplő adatokat (pl. egy letölthető Excel tábla vagy CSV fájl formájában);
- elérhetővé teszik a teljes kutatási adathalmazt, amire a cikk állításai épülnek;
- elősegítik a cikkben ismertetett adatok integrálását más publikációkban vagy weboldalakon található, szemantikusan kapcsolódó tudományos információkkal;
- géppel olvasható, nyílt formátumú metaadatokkal segítik a cikk megtalálását (pl. részletes bibliográfiai leírás a cikkről és a benne hivatkozott publikációkról, összefoglaló a cikk tartalmáról, fontosabb megállapításairól).

A cél az, hogy az online publikált cikkekben levő adatokat, információkat és tudáselemeket minél könnyebb legyen megtalálni, kiemelni, összekapcsolni és újrahasznosítani.

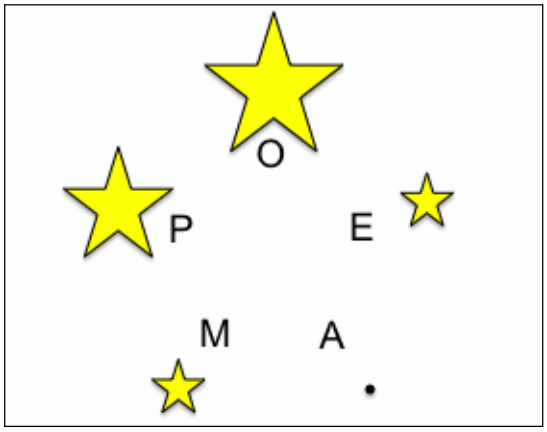
Ötcsillagos értékelés

Jelen cikk szerzője a bevezetőben említett, az Open Data kezdeményezésnél használt ötfokozatú jelzéshez hasonló rendszer bevezetését javasolja a szemantikus publikálás szintjének minősítésére, de nemcsak egyetlen skála mentén, hanem ötféle szempontból: lektorálás (P = Peer review), nyílt elérés (O = Open access), gazdagított tartalom (E = Enriched content), hozzáférhető adathalmazok (A = Available datasets) és géppel értelmezhető metaadatok (M = Machine-readable metadata). Az egyes fokozatokat 0-tól 4-ig terjedő számok vagy – vizuális formában – növekvő méretű csillagok jelzik *(1. ábra).*

Szakmai lektorálás

A tudományos cikkek szakmai bírálatának többféle formája, gyakorlata alakult ki az idők során: például van, ahol a bíráló(k) nevét csak a szerkesztő ismeri, máshol ezt közlik a szerzővel is, sőt esetleg az olvasók számára is publikussá teszik; a bírálók lehetnek felkért személyek, vagy bárki hozzászólhat, aki szakértő a témában (így alakulnak ki pl. az RFC-nek [Requests for Comments] hívott internetes szabványok); a véleményezés pedig történhet a publikálás előtt vagy azt követően (utóbbi esetben a szerző később közzétehet egy vagy több javított verziót). Mindegyik módszernek vannak előnyei és hátrányai: például a felkért lektorok gyakran leterheltek más, fontosabb munkákkal,

fennáll az összefonódás vagy éppen a visszaélés veszélye, utólagos bírálattal már nem lehet megakadályozni a gyenge vagy hibás publikációk megjelenését.



1. ábra. **Ennek a cikknek a minősítése a javasolt ötcsillagos rendszerben: P=3, O=4, E=1, A=0 *(ennél nem alkalmazható)*, M=1. Összesen: 9**

Az öt javasolt fokozat ezen a téren a következő:

0.–Nincs bírálat (a cikk előzetes szakmai lektorálás nélkül jelent meg, pl. az arXiv prepint archívumban).

1.–Publikáció előtti bírálat (a cikket véleményezte két vagy több felkért szakértő, és a megjegyzéseiket, illetve javaslataikat figyelembe véve fogadta el publikálásra a folyóirat szerkesztője; a lektorok nevét és véleményét titokban tartják).

2.–Bírálat, reagálási lehetőséggel (a cikket két vagy több szakember véleményezte – publikálás előtt vagy után – és a szerzőnek lehetősége volt reagálni a kritikájukra a szerkesztőnél, és/vagy módosításokat végezni a szövegen; a lektorok neve nem ismert és a véleményük nem kerül publikálásra).

3.–Publikáció utáni bírálat (a 2. szinten kívül a megjelent cikket utólag is kommentálhatják az olvasók, ezekről a szerző értesítést kap és reagálhat rájuk).

4.–Nyílt bírálat (a 3. szint kiegészítve azzal, hogy az egész folyamat teljesen transzparens, a felkért lektorok neve és véleménye is megjelenik a cikk mellett, melynek az eredeti és a javított verziói is hozzáférhetők).

Nyílt hozzáférés

Az *open access*, vagyis a tudományos eredményekhez való szabad hozzáférés az internet nélkül nem jöhetett volna létre. Ennek is több szintje és formája van, attól függően, hogy az egyes folyóirat-kiadók milyen üzleti modellt alakítottak ki. Hely szempontjából „zöld” szintűnek minősül az a publikáció, amely szabadon elhelyezhető például egy intézményi repozitóriumban is, míg az „arany” fokozat azt jelzi, hogy csak a folyóirat saját webhelyén érhető el nyilvánosan. Egy másik fajta osztályozás szerint ingyenes *(gratis)* az a cikk, amelynél csak a fizetési kötelezettséget oldották fel, de az elolvasás lehetőségén túl minden jogot továbbra is fenntartanak, ezzel szemben szabad *(libre)* az a publikáció, amelynél bizonyos felhasználási formákat is engedélyezett a jogtulajdonos (pl. valamelyik Creative Commons licenc formájában).

A javasolt kategóriák:

0.–Nincs nyílt hozzáférés (a cikk csak a folyóirat előfizetői számára érhető el; a jogokkal rendszerint a kiadó rendelkezik; önarchiválásra nincs engedélye a szerzőnek);

1.–Önarchiválás zöld/ingyenes formában (az előfizetéses folyóirat megengedi a szerzőnek, hogy repozitóriumba vagy egyéb helyre feltöltse, és így bárkinek hozzáférhetővé tegye a cikk valamelyik állapotát; a cikk bár szabadon olvasható, de nem használható fel más módon; a folyóirat honlapján pedig továbbra is csak az előfizetők férnek hozzá).

2.–Támogató által kért zöld/ingyenes hozzáférés (a cikkben ismertetett kutatást finanszírozó szerv – díjfizetés ellenében – engedélyt kap rá a kiadótól, hogy a cikk feltölthető legyen egy olyan ingyenes archívumba, mint a PubMed Central; a többi megkötés azonos az 1. szinttel).

3.–Szerző által finanszírozott arany/ingyenes hozzáférés (a szerző vagy munkahelye által befizetett díj fejében a kiadó a folyóirat honlapján ingyenesen olvashatóvá teszi a cikket; emellett esetleg az 1. szintű önarchiválásra is engedélyt ad a szerzőnek).

4.–Szerző által finanszírozott arany/szabad hozzáférés (a 3. szinthez hasonló, de a kiadó valamilyen Creative Commons vagy egyéb, szabad felhasználást nyújtó licenc alatt publikálja a cikket, azzal a feltétellel, hogy újrafelhasználás esetén hivatkozni kell a nála levő eredeti verzióra).

Gazdagított tartalom

A „Szemantikus publikálás” fejezetben ismertetett technológiák használatának szintjét jelzik ezek a kategóriák. Ilyen elemeket a szerzők is be tudnak építeni a publikációikba (linkeket helyezhetnek el

releváns oldalakra, vagy használhatnak Wordbe beépülő szemantikus szerkesztő modult [*ucsdbiolit.codeplex.com]* stb.).

0.–Nincsenek kiegészítések (az elektronikus verzió nem tartalmaz semmi pluszt egy nyomtatott kiadáshoz képest).
1.–Aktív hivatkozások (a cikkben kattintható linkek vannak más weboldalakra, adatbázisokra, illetve a hivatkozott publikációkra).
2.–Szemantikus elemekkel gazdagított szöveg (a szakkifejezések és fogalmak ki vannak emelve és pl. egy felugró ablakban más webes rendszerekből beemelt definíciók, képletek, adatbázislinkek jelennek meg, ha rájuk visszük az egeret; a hivatkozások típusa jelölve van).
3.–„Élő" tartalom (interaktív ábrák, szemantikus nagyítólencsék, amelyek megmutatják a grafikonok mögötti számsorokat, a hivatkozott cikkekből felbukkanó releváns részletek, átrendezhető irodalomjegyzék, és más hasonló interaktív elemek).
4.–Adatfúziók (*mash-up*-ok) (a cikkben ismertetett adatok más forrásokból származó információkkal vannak integrálva, pl. a földrajzi adatok letölthetők KML fájlként és Google-térképeken megjeleníthetők).

Adathalmazok hozzáférhetősége

Bár még a tudományos kiadók 2007-es Brüsszeli Deklarációja is hangsúlyozta a kutatási adatokhoz való szabad hozzáférés fontosságát, különösen a közpénzekből támogatott kutatások esetében, a tudósok egy része – érthető módon – vonakodik nyilvánossá tenni az esetenként sok fáradsággal összegyűjtött vagy előállított adathalmazokat, mielőtt alaposan ki nem elemezné őket és publikálná a belőlük levont következtetéseit. Az „adatok" alatt itt természetesen nemcsak számadatok értendők, hanem fotók, hang- és videofelvételek, grafikonok és egyéb ábrák, animációk és szimulációk, matematikai modellek, szoftverek és más hasonlók is. Sok szempontból nem szerencsés, ha a kiegészítő adatállományok csak a folyóirat webszerverén férhetők hozzá, érdemes ezeket intézményi repozitóriumokban, vagy még inkább az erre specializált szakterületi adattárházakban is elhelyezni a felhasználhatóság elősegítése és a hosszú távú megőrzés érdekében.

A javasolt szintek némi átfedést mutatnak a bevezetőben felsorolt, Tim Berners-Lee által ajánlott kategóriákkal:

0.– Nincsenek külön publikált adatok (csak a cikkben közölt adatokat ismerheti meg az olvasó; az ábrák és táblázatok nem tölthetők le külön; nincsenek további adathalmazok sem).
1.–Elérhetők kiegészítő fájlok (vannak a cikkhez kapcsolódó adatfájlok a folyóirat honlapján; a cikkben szereplő ábrák és táblázatok pedig külön is letölthetők, de nem újrahasznosítható formában – pl. csak TIFF vagy PNG képként).
2.–A cikk adatai letölthetők újrafelhasználható formában (a grafikonok és táblázatok adatai hozzáférhetők valamilyen szerkeszthető fájlban, pl. a számadatok táblázatkezelő vagy CSV formátumban).
3.–A háttéradatok is hozzáférhetők (a kutatás során keletkezett teljes adathalmaz publikálva lett valamilyen archívumban, egyedi URI vagy DOI azonosítóval, nyílt hozzáférést és szabad felhasználást biztosító licenc alatt, megfelelő részletességű metaadatokkal együtt, lehetővé téve így az adatok újraértelmezését és újrahasznosítását).
4.–Az adatokhoz hozzáfértek a lektorok (a publikálásra beküldött cikk mellett a felkért bírálók a teljes adathalmazt is tanulmányozhatták még a megjelenés előtt, így lehetőségük volt megítélni a cikkben levont következtetések helyességét).

Géppel olvasható metaadatok

Számos módszer létezik a digitális dokumentumok metaadatolására. Sok kiadó saját DTD-t definiált és használ a szerkesztési folyamat során a cikkek egyes részeinek (cím, szerzők, absztrakt stb.) jelöléséhez, de ezek a metaadatok a publikált PDF fájlból már rendszerint hiányoznak. A W3C-nél kidolgozott RDF és OWL2 szabványok lehetővé teszik egységes szótárak létrehozását ezeknek az információknak a kódolására, s így a különböző forrásokból származó metaadatok automatikus módszerekkel lekérdezhetők és egyesíthetők lesznek. A korábban már említett SPAR is egy ilyen ontológia, amely a tudományos publikációk leírására és szemantikus elemekkel való feldúsítására alkalmas. Különféle ajánlások léteznek az egyes szakterületeken arra vonatkozóan, hogy mi az a minimális leíró információ, amit bele kellene tenni a kutatási beszámolókba. Például a MIIDI (Minimal Information standard for reporting an Infectious Disease Investigation) szabvány a fertőző betegségekkel kapcsolatos publikációkhoz és adathalmazokhoz definiál egy speciális metaadat sémát, sőt egy MIIDI Editor-t is letölthetünk ezeknek a metaadatoknak a szerkesztéséhez (*http://www.miidi.org*).

A metaadatok tárolhatók akár magukban az XHTML, HTML5 vagy PDF fájlokban, akár külön XML, illetve RDF állományokban, vagy pedig olyan adatrepozitóriumokban, mint amilyen az Open Bibliography Project vagy az Open Citation Corpus. A publikáció lényegét összefoglaló metaadatok Open Research Reportként közzétehetők ilyenekre szakosodott „metaadat-folyóiratokban", míg a főbb tényszerű megállapítások önmagukban is közölhetők „nanopublikációk" formájában.

A javasolt minősítési szintek e téren:

0.–Nincsenek metaadatok (a cikk csak PDF fájlként jelent meg; a kiadó által a szerkesztési folyamat közben használt XML jelölések hiányoznak a publikált verzióból:

1.–Szerkezeti jelölés (a kiadó saját DTD-je által definiált tag-ek megtalálhatók a cikk XHTML változatában, vagyis jelölve van a cím, a szerzők névsora, az összefoglaló stb.).

2.–Bibliográfiai és hivatkozási metaadatok (a cikk teljes bibliográfiai leírása, valamint a hivatkozott publikációk metaadatai is letölthetők géppel olvasható fájlformátumban, vagy „Linked Open Data"-ként megtalálhatók egy hármas tárolóban *[triple store]*).

3.–Kibővített beágyazott jelölés (további szerkezeti, retorikai és szemantikus jelölők vannak a cikkben, RDF vagy más hasonló, számítógéppel értelmezhető módon kódolva).

4.–Strukturált cikkösszefoglaló (a cikkben szereplő fontosabb tények, hipotézisek, adatok és következtetések összefoglalása szabadon hozzáférhető mind emberi, mind pedig gépi fogyasztásra szánt formátumokban; az összefoglaló megfelel az adott szakterületen elvárt minimális információ ajánlásnak).

Bár a fentiekben ismertetett értékelési keretrendszer ötlete még nagyon új és a gyakorlati alkalmazásához minden bizonnyal finomításokra lesz szükség, a Ubiquity Press máris jelezte, hogy használni szeretné az általa publikált cikkek minősítéséhez ezt az ötcsillagos szisztémát.

/SHOTTON, David: The Five Stars of Online Journal Articles – a Framework for Article Evaluation. = D-Lib Magazine, 18. köt. 1–2. sz. 2012./

(Drótos László)