
# A *linked data* és a *big data* találkozása – a tudásszervezési rendszerek szempontjából

## Big data

A változatos, komplex és hatalmas méretű digitális adathalmazok megjelenése a weben életre hívta a *big data* nevű jelenséget. Ilyen „nagy adatokat" termel például a közösségi média, az elektronikus kereskedelem, a kormányzat, a tudományos kutatás... A fogalomnak sokféle meghatározása született; másként közelítik meg ezt a témát a tudósok, a számítástechnikával és az információtudománnyal foglalkozó szakemberek, a tudománypolitikát irányítók és a finanszírozó szervek vezetői. Vannak, akik azt a technológiai kihívást hangsúlyozzák, amit az igen nagy méretű adatállományok – beleértve természetesen a számadatokon kívül a szöveges és a multimédia tartalmakat is – okoznak, és ezért a *big data* kifejezés alá veszik mindazokat az eszközöket és eljárásokat is, amelyekkel létrehozni, kezelni, feldolgozni és tárolni lehet ezeket az óriási adathalmazokat. Olyan korszerű technológiákról van itt szó, mint például a hagyományos relációs adatmodellt meghaladó noSQL adatbázisrendszerek, a klaszterbe vagy gridbe kötött gépeken való párhuzamos adatfeldolgozásra kitalált MapReduce programozási elv, a természetes nyelvfeldolgozás, a gépi tanulás, vagy az újfajta vizualizációs megoldások.

A Digging into Data nevű kezdeményezés (*diggingintodata.org*), melyet több ország kutatásfejlesztési alapjai közösen indítottak, a humán és a társadalomtudományok területén szerveződő *big data* projekteket támogatja. Mivel a világunk egyre digitálisabbá válik és egyre nagyobbra nőnek az adatbázisok (a digitalizált könyvektől, újságoktól és zenéktől kezdve az olyan tranzakciós adatokig, mint a webes keresések naplói, a szenzorok mérései, vagy a mobil hálózatok cellainformációi), ezért új megoldásokra, újfajta kutatási infrastruktúrára van szükség a digitális adatok kereséséhez, elemzéséhez és megértéséhez.

*Simon Hodson*, a brit akadémiai szféra infokommunikációs infrastruktúráját működtető *JISC* szervezet kutatási igazgatója szerint számukra ezeken a területeken jelent kihívást a *big data* jelenség: webarchiválás, tanuláselemzés, felhasználói statisztikák és kutatási adatok. A JISC szponzorálja az *Oxford Internet Institute* egyik projektjét, amely a *British Library* kezelésében lévő webarchívum társadalomtudományi célú hasznosíthatóságát kívánja demonstrálni: az 1996 és 2010 között az *.uk* domén alá tartozó szerverekről begyűjtött mintegy 30 terabájtnyi weboldalból kivonják a hiperlink gráfokat, hogy statisztikai elemzéseket végezzenek rajtuk.

Az amerikai *National Science Foundation* és a *National Institutes of Health* BIGDATA nevű kutatási programjának felhívásában a *big data* fogalmát így határozták meg: nagy, változatos, komplex, longitudinális (vagyis idősorszerűen rögzített) és/vagy elosztott adathalmazok, melyek forrásai különféle eszközök, szenzorok, internetes tranzakciók, e-mail-, video- és kattintássorozatok, vagy egyéb digitális források. A program olyan tudományos és műszaki megoldásokat támogat, amelyekkel ezeket az adathalmazokat kezelni, elemezni, megjeleníteni lehet, és társadalmi, gazdasági, egészségügyi, életminőségi stb. célok érdekében hasznos információkat lehet kinyerni belőlük.

Bár maga a fogalom elég tisztázatlan még, abban egyetértés van a témában publikálók között, hogy az ún. *linked data* vagy *open data* kategóriába tartozó, vagyis a strukturált és többnyire nyilvánosan hozzáférhető adathalmazok is részei a *big data* világának, sőt: a szemantikus web alapját jelentő, automatizáltan értelmezhető és egymással kombinálható linkelt adatok ideális tesztkörnyezetet adhatnak a *big data* kutatások számára. Akár strukturált, akár strukturálatlan halmazokról van szó, hasonló műszaki kihívásokkal kell megbirkózni ezek rendszerezése, karbantartása, menedzse-

lése, megőrzése, feltárása, vizualizációja, hozzáférhetővé tétele és használata során.

## Linked data

A formalizált, strukturált és rendszerezett adatok a *big data* egyik típusát jelentik. A *linked data* és annak olyan speciális alkalmazásai, mint a linkelt kötött/korlátozott szótárak és a tudásszervezési rendszerek, szilárd szemantikus alapként szolgálhatnak a rendezetlen adatok osztályozásához és ábrázolásához. Felhasználhatók például automatikus vagy félautomata szövegelemzésekhez, tematikus metaadatoláshoz, és az adatok facettás, kategorizált vagy hierarchikus megjelenítéséhez.

Nagy tömegű szöveges adat értelmezése esetén olyan technikákra van szükség, mint amilyen a szemantikus szövegelemzés, a természetes nyelvi feldolgozás, az adatbányászat és az adatvizualizálás. A *W3 Konzorcium* által fejlesztett *SKOS* (*Simple Knowledge Organization System*) specifikáció egyfajta összekötő hídként szolgál a különféle tudásszervezési formák (tezauruszok, osztályozási rendszerek, tárgyszórendszerek, taxonómiák és folkszonómiák) valamint a *linked data* közösség között. A SKOS-alapú, linkelt, kötött szótárak megfelelő szemantikus keretet nyújtanak a *big data* halmazok elemzéséhez és ábrázolásához is. A SKOS-nak köszönhetően a különböző tezauruszok és egyéb szótárak leképezhetők egymásra és összekapcsolhatók, s így keresztül-kasul kereshetővé és böngészhetővé válnak a linkelt adatokat tartalmazó repozitóriumok, a nyílt archívumok, a digitális könyvtárak, valamint a különféle keresőrendszerek és szolgáltatások. Már vannak olyan nagy szótárak, amelyeket SKOS-formátumba is átkódoltak és linkelt adatforrásokként felhasználhatók. Ilyen például a környezettudományi GEMET, az agráripari AGROVOC, az orvostudományi MESH, az interdiszciplináris szótárak építését segítő HIVE, a *Kongresszusi Könyvtár* LCSH tárgyszórendszere, a gazdaságtudományi STW Thesaurus for Economics és az oktatási területeken használatos ScOT.[1]

A *Bernard Vatant* és *Pierre-Yves Vandenbussche* által létrehozott Linked Open Vocabularies honlap (*lov.okfn.org*) a linkelt nyílt szótárak hasznos nyilvántartása. Ezek nemcsak a jól strukturált repozitóriumok és szemantikus webes alkalmazások számára érdekesek, hanem felhasználhatók rendszerezetlen szöveges adatok indexelésére, rendezésére és analizálásra is.

Olyan webes szolgáltatások is léteznek már, amelyek az RDF adatleíró keretrendszerre és a *linked data* szabványokra alapozva újszerű megoldásokat kínálnak a *big data* jelenséggel járó kihívásokra. A Zemanta (*zemanta.com*) böngészőkiegészítő például releváns címkéket, valamint külső forrásokból linkeket és képeket javasol automatikusan blogbejegyzések vagy cikkek, hírek írása közben. A Calais (*opencalais.com*) nevű szolgáltatással strukturálatlan szövegekből készíthetünk RDF formátumú kimenetet. Használható blogbejegyzések címkézéséhez, de akár múzeumi gyűjtemények kategorizálásához is. Jó példa a SKOS-alapú tezauruszokra a PoolParty (*poolparty.biz*), amely képes összekötni a különböző, linkelt adatokat tartalmazó repozitóriumokat, megkönnyítve így a keresést és az információgyűjtést.

## Példák

– Agráripari információk menedzselése és megosztása: A *FAO* (*Food and Agriculture Organization*) leíró metaadatok, tezauruszok, AGROVOC szótárak és ontológiák felhasználásával sokféle formátumban gyűjti, strukturálja és terjeszti az éhínség elleni harchoz szükséges táplálkozási, élelmiszerügyi és mezőgazdasági információkat.
– Személyre szabott egészségügy: Betegek adatait orvostudományi ontológiák segítségével nagy tömegben kielemezve és az adatok között mintázatokat keresve, az orvosok jobb döntéseket hozhatnak és személyre szabott betegségkockázati profilokat állíthatnak fel.
– Humán témák: A *DPLA* (*Digital Public Library of America*) és az Europeana egyaránt jó *big data* példa a művészetek és a bölcsésztudományok területén. Ezek a szervezetek egyrészt adatforrásként szolgálnak (az általuk biztosított API szolgáltatások révén), másrészt nagy adatfeldolgozó szervezetek is. Gondosan és folyamatosan fejlesztik az adatmodelljüket, és fontosnak tartják a szemantikailag gazdag metaadatokat, mert ezeknek köszönhetően a felhasználóik intuitívabb, intelligensebb módokon tudnak keresgélni a gyűjteményeikben.

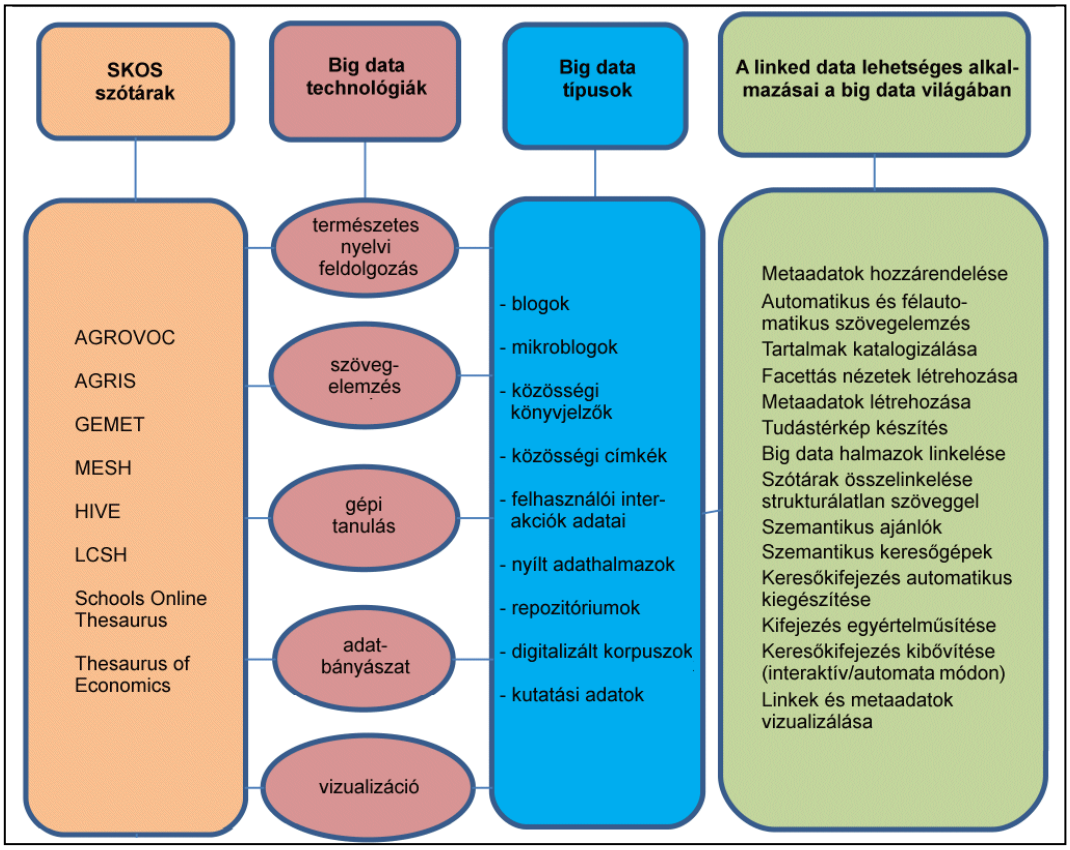
[1] A http://www.w3.org/2001/sw/wiki/SKOS/Datasets oldalon felsorolt források közt ott találjuk az OSZK tezauruszait is. (A ref.)

## Alkalmazási területek

A SKOS-alapú nyílt és linkelt szótárak többek között az alábbi területeken lehetnek hasznosak a *big data* típusú adathalmazok feldolgozása és használata során:

– Egyidejű keresés és böngészés nyílt adatrepozitóriumokban és *big data* forrásokban.
– Linkelt szótárak összekapcsolása strukturált, részben strukturált és strukturálatlan adatokkal.
– Nagy szöveghalmazok általános jellegű, illetve szakterületre specializált, természetes nyelvű feldolgozása.
– Tematikus metaadatok vizuális ábrázolása.
– Szemantikus keresőgépek és ajánlórendszerek kifejlesztése.
– Keresőfogalmak és -kifejezések automatikus kiegészítése.
– Keresőkérdések interaktív és automatikus kibővítése.
– Nézetek szerinti (facettás), illetve kategóriák alapján való böngészési lehetőség megteremtése.
– Digitális szövegek elemzése és feldolgozása humántudományi projektekben.

Néhány további lehetséges alkalmazás látható még az 1. ábrán.




1. ábra **Potenciális *linked data* alkalmazások a *big data* világban**

**/SHIRI, Ali: Linked data meets big data: a knowledge organization systems perspective = Advances in Classification Research Online, 24. évf. 1. sz. 2014. p. 16-20/.**

**http://journals.lib.washington.edu/index.php/acro/article/view/14672**

*(Drótos László)*