Reflektorfényben a kutatási adatok: adatkatalógus a megosztás és újrafelhasználás támogatására

Az utóbbi években több vizsgálat során is bebizonyosodott, hogy a szakcikkekben közölt eredmények megismételhetősége kétséges. Több területen (például közgazdaságtudomány, pszichológia) a vizsgált minták alig egyharmada esetében sikerült maradéktalanul reprodukálni a közölt eredményeket. Ugyanakkor egyes publikációs házak, mint például a *PLOS* (*Public Library of Science*), vagy olyan meghatározó irodák, mint amilyen a *White House's Office of Science and Technology Policy* (a Fehér Ház Tudomány- és Technológiapolitikai Irodája, USA) már megkövetelik, hogy a kutatók megosszák adataikat vagy legalábbis készítsenek adatmegosztási tervet. A könyvtáraknak kapóra jön ez az új feladat: meglévő szakmai tudásukat és tapasztalataikat jól kamatoztathatják e feladatok vállalásánál.

A szerzők a *New York University* egészségtudományi könyvtára munkatársaiként tapasztalhatták meg, hogy a kutatóknak gyakran okoz főfájást a kutatási anyagok elhelyezése. Ha a tudós például rá is szánja magát valamilyen elérhető népegészségügyi adathalmaz felhasználására, rengeteg technikai akadályba botlik. Az adathalmazokhoz vezető hozzáférési útvonalak olykor kifürkészhetetlen labirintusnak tűnnek: az adatok szét vannak szabdalva és szét vannak szórva, több weboldalról érhetők csak el. Megtévesztők lehetnek a licencelési feltételek, hiányosak a kereséshez szükséges szótárak. Hasonló nehézségek hívták életre a New York-i Egyetemen azt a programot, amely összehangolta az intézményi egészségtudományi könyvtár osztályozási rendszerét és webfejlesztési kapacitását a népegészségügyi kutatók tapasztalataival. A projekt célja az volt, hogy segítsék a kutatókat a megfelelő adathalmazok fellelésében, saját adataikat pedig hasznosítható módon el tudják helyezni a weben.

E célok teljesítéséhez speciális metaadatséma fejlesztésére volt szükség. A házilag tervezett adatkatalógus gazdag metaadatséma elemeire épít, sok funkcióval felszerelt, fazettás kereséssel is kiegészített weboldalas felületet biztosít. Ugyanakkor a nem technikai személyzet számára is kialakítottak egy olyan adminisztrációs felületet, amelyen keresztül segítség nélkül is kezelni lehet az ún. CRUD-műveleteket (Create, Read, Update, Delete). Egyes kutatásokhoz hatalmas külső adatállományokat használnak – ilyenek például az amerikai népszámlálási adatok. A rendszerben költségkímélő megoldással ezeket nem tárolják helyileg, így nincs szükség hatalmas adattároló kapacitásra sem.

Alapozás

Az adathalmazok kezelőrendszerének fejlesztéséhez a könyvtárosok első lépésként számbavették a használatban lévő adatkatalógusokban fellelhető metaadatsémákat. Remek példája a megosztott kutatási adatoknak a 2014-től a *Nature* folyóiratakéként megjelenő *Scientific Data* felülete, ahol részletes leírásokat is közzétesznek a tudományos adatokról. Megvizsgálták továbbá a *Dryad*, a *DataCite* és a *W3C* adatkatalógus metaadatszótárát, a *National Institution of Helath* (NIH, Nemzeti Egészségügyi Intézet, USA) majd félszáz repozitóriumának sémáját, ezek alapján sikerült meghatározni a leggyakoribb metaadatelemeket. Figyelembe vették a *NIH BioCADDIE* kutatásiadat-szabványjavaslatát is, amely általános, az összes tudományterületen alkalmazható sémát alakított ki.

A fentieken kívül számos további, sikeresen működő adatkatalógus szerkezetét tanulmányozták át. Több helyen az adathalmazok kezelésére a *Drupal* rendszert használják. Első pillantásra, minthogy a fejlesztő intézmény több webhelye is ezzel az eszközzel készült, ez a platform megfelelőnek tűnt az adathalmazokban való eligazodás céljára is, ám közelebbi megvilágításban számos keresési nehézség merült fel, az ötletet emiatt

elvetették. Több hasznos tanulsággal járt a *National Snow & Ice Data Center* (Nemzeti Hó- és Jégadat-Központ, http://nsidc.org/data/search/) rendszerének elemzése is.

Egy teljesen új világ

A prototípus *Solr, Backbone.js, jQuery and Underscore.js* alkalmazásokra épült. Sikerült beépíteni a publikálási és eseménykezelő rendszert. A Backbone csupán keretrendszert tudott biztosítani, viszont számos további funkció beágyazására is szükség volt, ilyenek például a beléptetés, a tartalomkezelés vagy a komplex keresés. Az összetettebb, kiegészítő feladatokra ezek az alkalmazások tehát már nem bizonyultak megfelelőnek. Ezért hagyományos szerveroldali fejlesztésben kezdtek el gondolkodni, a következő kísérletre a *Symfony*-t (https://symfony.com/) szemelték ki. A Symfony beléptetőjét sikerült összehangolni az intézet authentikációs adataival.

Az adathalmazok metaadatsémái rendkívül változatosak, a leírandó objektumoktól, adattípusoktól függően más-más elemkere van szükség. Külön eljárás a változatos forrásból származó és más-más szerkezetű adatok „lefordítása” normalizált relációs adatbázis számára, az adattáblák kialakítása, az adatok közötti kapcsolatok definiálása. Meghatározták az általános érvényű adattípusokat (például „Kiadó”, „Szakember” – ez utóbbi az adathalmaz szakértője, akihez segítségért lehet fordulni), valamint definiálták a speciális, egy-egy adathalmazon belül érvényes elemek körét is. A végső modell 24 entitást és 54 adatbázistáblát tartalmaz. Egy-egy teljes adathalmazrekord megjelenítéséhez a Symfony akár 20-30 adatbázis-lekérdezést is elvégez. A rendszer teljesítményét jól segíti a *Doctrine Object-Relational Mapping* (ORM) eszköz (http://www.doctrine-project.org/), amely a háttérben megfelelően kezeli a legtöbb adatbázis-kommunikációt. A definiált osztályok, tulajdonságok és kapcsolatok beállításait az alkalmazott háttértámogató rendszerek intelligensen és gyorsan kezelik, az eredmény kielégítőnek bizonyult.

Adminisztrációs felület

Az utolsó feladat az adminisztrációs felület kialakítása volt. A komplex adatmodell és a különféle metaadatelemek, amelyek a külső és a belső adathalmazok leírásához és kezeléséhez szükségesek, valamint további tizenhat kapcsolódó entitás (például „Kiadó”, „Helyi szakértő”, „Pályázat/Támogatás”) beágyazása viszont már olyan mértékű többletet jelentett, amely a Symfony rendszer képességeinek határait feszegette. Nehézségek merültek fel a sokféle metaadat átlátható, űrlapos megjelenítésében és a besorolási adatállományok kezelésében is. A könyvtárosok egyformán igényelték az adatbeviteli felületek mezőinél a begépeléskor aktivizálódó automatikus kiegészítő funkció telepítését és a törzsadatok munkamenet közbeni szerkeszthetőségét, illetve ugyanitt új rekord létrehozását, anélkül, hogy el kelljen hagyniuk a megkezdett kezelőablakot. A kétféle bejegyzési módszer integrálását a Symphony rendszerében, némi nehézségekkel ugyan, de sikerült megvalósítani.

Némi fejtörést okozott az űrlap tagolásának kialakítása: legyen-e rajta az összes beviteli mező egyetlen hosszú űrlapon, vagy inkább tagolódjon az adatbeviteli felület több oldalra, külön-külön fülekkel megnyitható adatcsoportok szerint? Hosszú megfontolás után az egyetlen, hosszú adatlap megjelenítése mellett döntöttek, a több oldalra szabdalt, fülek mögé rejtett részinformációk ugyanis ellehetetlenítették volna a teljes adatsor áttekintését, a más szakaszokkal való összefüggések megtalálását, adott esetben nehezítették volna a felesleges szakaszok átugrását is. Az űrlap takarásban lévő részei az adatelemek felviteli pontjainak kezelését is hátráltatták – ilyen eset például az a szituáció, amikor felmerül a kérdés, hogy vajon a „Helyi szakértő” mező történetesen a *Kapcsolatok* oldalon vagy a *Hozzáférési információk* léptetőgomb mögött rejtőzik-e. Az egyoldalas hosszú űrlapon ugyanakkor az adattípusok csoportjai jól átláthatók a panelek fejléceinek a segítségével.

A végeredmény

A *Data Catalog* 2015-ben kelt életre, fejlesztését saját intézményi igények indították el. Az *Apache Solr* keresőre és a Symfony2 PHP keretrendszerre épülő, saját megoldásokat ötvöző rendszer már az első hónapokban jelentős látogatottságra tett szert. Az új platform segíti a saját kutatásból eredő anyagok elhelyezését és megosztását, valamint a külső adatforrások felkutatását és hasznosítását is, egyúttal támogatja a tudományos kutatások előmenetelét és minőségét. A következő fejlesztési lépés a forráskód GitHub-on való közzététele, hogy más intézmények – immár sokkal kevesebb időráfordítással – hasonló katalógust telepíthessenek. Az alkalmazás JSON és Linked Open Data (Nyílt kapcsolt adatok) kimenetet biztosít JSON-LD formában, valamennyi oldalra beágyazódóan. Ezzel a katalógus összeköthetői a hasonló tárházakat,

de emellett biztosíthatja az intézmények kutatási anyagai közötti metakeresést is. A Data Catalog címe: http://datacatalog.med.nyu.edu.

Irodalom

READ, K. – ATHENS, J. – LAMB, I. [et al.]: Promoting Data Reuse and Collaboration at an Academic Medical Center. International Journal of Digital Curation, 10. évf. 1. sz. 2015, 260–267.
http://doi.org/10.2218/ijdc.v10i1.366

MEMBERS, W.: WG3-Metadata Specifications: NIH BD2K bioCADDIE Data Discovery Index WG3 Metadata Specification v1. 2015. Zenodo.
https://zenodo.org/record/28019#.VsdMMPkrJhE

/LAMB, Ian – LARSON, Catherine: Shining a light on scientific data: Building a data catalog to foster data sharing and reuse. = Code4Lib Magazine, 69. köt. 32. sz. 2016.
http://journal.code4lib.org/articles/11421 /

(Dudás Anikó)