

Informatika a felsőoktatásban 2014 konferencia, Debrecen, 2014. augusztus 27–29.

Tudományos repozitóriumok az MTA-ban: a KDK¹ és a SZTAKI² tanulságai

Scientific repositories at the Hungarian Academy of Sciences: the Research Documentation Centre and SZTAKI

Micsik András^a, Gárdos Judit^b

^aMTA SZTAKI

micsik@sztaki.mta.hu

^bMTA Társadalomkutató Központ

Gardos.Judit@tk.mta.hu

Absztrakt: Két olyan repozitóriumot mutatunk be a hazai palettáról, melyeknek elsődleges célja a kutatás szolgálata. Ez a feladat új követelményeket támaszt a repozitóriumokkal szemben, kezdve az örvedetesen terjedő nyílt hozzáféréstől a kutatási adatok kezeléséig. A kutatások közzététele ma már nem korlátozódhat a közlemények megjelentetésére, a kutatási adatok és a módszerek, algoritmusok nyilvánosságát is meg kell teremteni. Egy tudományos repozitóriumban lehetővé kell tenni minél színesebb, sokfélebb háttéranyag tárolását, ezzel is elősegítve a tudományos eredmények terjesztését, újrafelhasználását, reprodukálását és ezáltal ellenőrizhetőségét. A tudományágak szokásai határozzák meg, hogy a másodelemzés mennyire elfogadott és elvárt, mely kutatási dokumentumok válnak értékessé és megőrizendővé. Az MTA irányelvei az utóbbi időben erőteljesen változtak e tekintetben. A kutatási adatok az általános problémákon túl, mint például hivatkozhatóság, tudományáganként speciális követelményeket is támasztanak. A szociológia esetében az adatok között találhatunk videókat, vagy kérdőíves válaszokat. Mind a két esetben felmerül a kérdés, hogyan tudná ezeket egy külföldi kutató használni? A tudományos adatok létrehozásának kontextusa fontos értelmezési szempont a másodelemzés során, ennek érdekében tudományágra szabott, extenzív metaadat tárolási igényeket kell kiszolgálni. A repozitóriumok a kutatástámogatás és kutatómenedzselés számára is nyújthatnak értékes adatokat. Ahhoz, hogy a repozitóriumunk mindezeket lehetővé tegye, legtöbbször az informatikai megoldás erőteljes konfigurálása és testre szabása szükséges. A társadalomtudományos adatok szenzitivitása, az anonimizálási lehetőségek és nehézségek korlátozhatják a nyílt hozzáférhetőséget. Két gyakorlati megvalósítás kapcsán nyújtunk ízelítőt a felmerült problémákból, azok megoldásaiból, valamint kitérünk a világ élvonalbeli trendjeire, fejleményeire is e téren.

Kulcsszavak: repozitórium, digitális archívum, kutatási adat

Abstract: The paper presents two recently created repositories established for supporting researchers and scientific dissemination. Scientific publication today requires solutions beyond publishing papers for various other scientific data management issues such as the access, re-use and citation of raw data. This vastly increases the formats and types of items stored in repositories. Raw data for example produced in sociology may need detailed and translated documentation of its internal structure. Repositories may also be an important data source for funding and research management. In this paper we describe the process of creating two scientific repositories in the fields of computer science and social sciences, and provide examples of the problems and solutions we faced during the implementation.

Keywords: scientific repositories, digital archives, research data

¹ MTA TK Kutatási Dokumentációs Központ, <http://www.openarchive.tk.mta.hu>

² SZTAKI Repozitórium, <http://eprints.sztaki.hu>

1. Bevezetés

A kutatás szolgálata új követelményeket támaszt a repozitóriumokkal szemben, kezdve az örvendetesen terjedő nyílt hozzáféréstől a kutatási adatok kezeléséig. Ezek a feladatok konkrét kutatási témákként jelentkeznek az informatika, a tudománymenedzsment, sőt a tudományelmélet területén manapság. A témák között a nyers adatok közzététele, azok hivatkozása, a szerzők egyértelmű azonosítása, de a digitális archívumok működését vizsgáló elméleti kérdések is megjelennek [1]. A fejlődés elég gyors, példaként említhető az ORCID [2], amely egy globális szerző azonosító, vagy a DataCite [3] hálózat, amely a kutatási adatok megőrzését és publikálását tűzte ki célul. A kutatási adatok archiválásának általános kérdésein túl tudományáganként speciális követelmények is megjelennek. A szociológia esetében az adatok között található videókat, vagy kérdőíves válaszokat. Mind a két esetben felmerül a kérdés, hogyan tudná ezeket egy, a kutatás részleteit nem ismerő, vagy akár egy külföldi kutató használni. Az informatikatudomány esetében az eredmények ellenőrizhetősége, reprodukálhatósága egyre fontosabb, és ehhez kapcsolódóan nem csak a nyers adatok, hanem az algoritmusok, szoftverek publikálása is szükségessé vált.

Európában, de világszerte is egyre-másra szerveződnek társadalomtudományos archívumok. Ezek sokszor intézményi vagy nemzeti archívumok. Talán a legnagyobb európai, társadalomtudományos adatokat kezelő szervezet a CESSDA.³ Világszinten a legnagyobb társadalomtudományos archívumokat tömörítő szervezet az IASSIST. Mindkét intézmény fő profilja a *kvantitatív* módszerrel készült adatok tárolása (tipikusan kérdőíves felvételek adatbázisai). A társadalomtudományos, *kvalitatív* adatokkal foglalkozó archívumoknak is van egy kis európai szervezete.⁴

Az Európai Uniónak sok éve célja, hogy úgynevezett kutatási infrastruktúrákat (Research Infrastructure) alakítson ki. Ennek keretében több európai bölcsészet- és társadalomtudományos archívumot támogatott. Az utóbbi néhány évben ezeket az archívumokat egységes rendszerbe kívánja terelni, két éve fut a DASISH⁵ projekt, amely az öt eddig támogatott európai bölcsészet- és társadalomtudományos archívum közötti szinergiákat, együttműködési lehetőségeket és lehetséges közös standardokat tárja fel. Magyarországon 2014 tavaszán kapott újra életre a NEKIFUT rendszer, amely a magyarországi kutatási infrastruktúrákat térképezi fel (ideértve a társadalomtudományosokat is).⁶

2. A SZTAKI Repozitórium

Az MTA SZTAKI 2013 elején létrehozta a saját repozitóriumát azzal a céllal, hogy az intézetben keletkező minden olyan a tudományos kutatás és fejlesztés, valamint műszaki innováció során keletkezett digitális anyagot tároljon, mely valamilyen megőrzendő értéket tartalmaz. Az archívum elsődleges gyűjtőköre tudományágak szerint: műszaki tudományok és technológiák, számítástechnika, számítógép tudomány, irányítás és vezérlés. A SZTAKI Repozitórium nem csak a tudományos közlemények teljes szövegének tárolására ad lehetőséget, hanem a kapcsolódó anyagok, például prezentációk, nyers adatok, videók, stb. elhelyezésére is. A tudományos anyagok mellett műszaki jelentések, tanulmányok, segédletek, stb. is feltölthetők. A repozitóriumban az MTA SZTAKI munkatársai

³ <http://www.cessda.net>

⁴ EQUALAN, <http://www.iqda.ie/content/equalan>

⁵ <http://www.dasish.eu>

⁶ <http://www.nih.gov.hu/strategiaalkotas/kfi-infrastruktura/kutatasi-infrastrukturak>

helyezhetnek el anyagokat, és azok elérhetőségét három szinten korlátozhatják: mindenki, az intézet munkatársai, a repozitórium adminisztrátorai.

A fenti célok megvalósítására a rugalmasan konfigurálható EPrints rendszert választottuk, mely a University of Southampton által immár 13 éve fejlesztett szabad szoftver. Az elmúlt másfél év alatt jó tapasztalatokat szereztünk a rendszerről, mivel jelentős felhasználói közössége van, kicsi az erőforrásigénye (gyengébb hardver konfigurációban is gyors válaszidővel működik), és hasznos kiegészítő bővítményekkel rendelkezik.

Az EPrints konfigurálását az autentikációval kezdtük: a SZTAKI LDAP szerverét használjuk a felhasználó azonosítására, az EPrints wikiben található recept alapján. Az autentikáció során a felhasználó LDAP adatainak egy részét áttöltjük az EPrints-be, így például a név vagy a részleg változásai automatikusan átkerülnek az EPrints-be. Ilyenkor megkapjuk a felhasználó LDAP azonosítóját is, és jobb híján ezt kezdtük el használni a szerzők egyértelműsítésére is. Később ezt kiegészítettük a szerző ORCID és MTMT azonosítóival is. Ezen szerző-azonosítás alapján megoldottuk, hogy a bejelentkezett felhasználó az általa feltöltött tételeket tudja módosítani, amely egyébként nem megoldott az EPrints-ben, mivel a szerzők és a felhasználók két külön rekordkészletet alkotnak. Az új tételeket feltöltés után a szerkesztők hagyják jóvá, a bevitt metaadatok ellenőrzése után.

Az EPrints tételek (ún. eprint-ek) metaadat mezőit könnyedén lehet módosítani, így felvettünk több olyan mezőt is, amelyek a pályázati támogatást rögzítik, illetve a weblapunkon való megjelenést segítik (pl. magyar nyelvű kulcsszavak). Továbbá az EPrints-ben szabályozni tudjuk, hogy mely mezők legyenek kereshetőek, illetve mely mezők alapján készüljenek automatikus listázások, és azok hogyan jelenjenek meg. Erre példaként készítettünk olyan listázást, amely a szerző LDAP azonosítója alapján gyűjti ki a szerző publikációit.

A repozitóriumot feltöltöttük a korábban más rendszerben tárolt publikációs metaadatokkal, ezzel lehetőséget adva, hogy kollégáink utólagosan egyszerűen feltölthessék cikkeik teljes szövegét. A 2013-tól életbe lépett MTA open access rendelet [4] kötelezővé tette a közleményeink teljes szövegének elérhetővé tételét, amelyet már a SZTAKI Repozitóriummal is tudunk teljesíteni, mivel az megfelelt az MTA repozitórium minősítési folyamatban. A folyamat során definiálnunk kellett a repozitórium működésével kapcsolatos keretfeltételeket, amely egy hasznos ellenőrző listaként szolgált számunkra is. Ennek során a repozitórium működtetésének személyi és műszaki feltételeit kellett tisztázni, valamint a tartalommal kapcsolatos minőségi jellemzőket felmérni. Ezek között említendő elemek: a hardver üzemeltetésének, a mentések rendszerességének biztosítása, a bekerülő anyagok ellenőrzési folyamata, és a más szerverekkel való kapcsolatok.

A repozitóriumok hasznossága nagymértékben a külső láthatóságon, kapcsolatokon múlik. A legáltalánosabban használható kapcsolódási mód az OAI-PMH protokoll [5], ez alapértelmezésben be van kapcsolva az EPrints-ben. Az OAI kapcsolódási pontot a ROAR [6] és OpenDOAR [7] nyilvántartásokban jelentettük be.

Az OpenAIRE project [8] az EU FP7 és H2020 által támogatott közleményeket gyűjti. Az ő nyilvántartásukba is lehet csatlakozni, ennek feltétele az, hogy az általuk megadott módon tároljuk a finanszírozási adatokat, és az FP7 illetve H2020 által támogatott tételek egy OAI set-ben kigyűjthetőek legyenek. Mindezeket egyetlen EPrints bővítmény telepítésével meg lehet oldani.

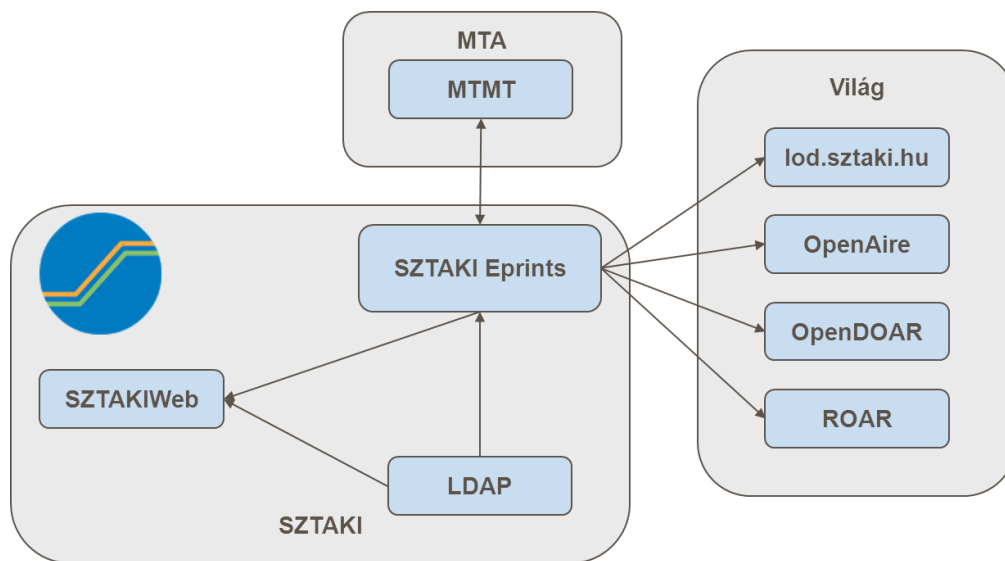
Egy másik kapcsolódási pontunk az MTMT-ből teszi lehetővé a PDF fájlok közvetlen feltöltését a közlemény adatainak megadásával egyidejűleg. Ehhez a SWORD protokollt használjuk, melyet az EPrints-ben könnyen be lehet kapcsolni, és minősített repozitóriumok esetén az MTMT rövid tesztelés után engedélyezi a SWORD kapcsolatot. A SWORD

kapcsolaton keresztül feltöltött közlemények a szerkesztők postaládájában jelennek meg, a főbb metaadatokat kitöltve látjuk már, de ellenőrzésre szorulnak. Fontos, hogy az ilyen tételek esetén tároljuk a közlemény MTMT azonosítóját is, ebből láthatjuk, hogy a tétel az MTMT-ben már rögzítve van.

Sok belső vitánk volt arról, hogy a publikációkat először az MTMT-be töltsük fel, vagy a saját repozitóriumunkba. Univerzális megoldás nem született, ezért mindkét irányt támogatjuk, az egyik irányban a SWORD kapcsolaton keresztül, míg a másik irányban az EPrints RIS export lehetőségét igazítottuk az MTMT RIS import funkciójához. Ez utóbbit akkor érdemes használni, ha még a közlemény megjelenése előtt közzé akarjuk tenni eredményünket preprint-ként, és ilyenkor a megjelenés adatait (oldalszám, DOI, stb.) még nem tudnánk az MTMT-ben kitölteni.

Az MTMT és a SZTAKI repozitórium szinkronban tartására írtunk egy szkriptet, amely az MTMT-ből exportált XML alapján kiegészíti az Eprints-ben tárolt tételeket.

A repozitórium adatait RDF-be konvertálva áttöltjük saját Linked Open Data szolgáltatásunkba is (lod.sztaki.hu). Végezetül, a SZTAKI intézeti portáljába több helyen be vannak integrálva az EPrints szerverről letöltött adatok. A személyek oldalain és a részlegek oldalain is megjelennek a kapcsolódó publikációk, melyek többféleképpen rendezhetők is. Ezen kívül kulcsszavak és témakörök szerint is kereshetők a publikációk. Mindezek az adatok a repozitórium felé továbbított egyedi keresések eredményeként kerülnek be a portál weboldalaiba. Az ismertetett kapcsolatokat összegzi az 1. ábra.



1. ábra. A SZTAKI Repozitórium kapcsolatrendszere más szolgáltatásokkal

3. A KDK

Az MTA Társadalomkutató Központjának (TK) Kutatási Dokumentációs Központja (KDK) az MTA központi határozata után született meg 2013 februárjában. A KDK szakmai vezetője (Kovács Éva) és munkatársai a 20. Század Hangja Archivum és Kutatóműhely⁷ alapítóiként több éves társadalomtudományos archívumi tapasztalattal rendelkeznek. A 20. Század Hangja

⁷ <http://www.20szadhangja.hu>

Archívum az Open Society Archive-val (OSA) együttműködésben, az OSA szerverén⁸, a DSpace nyílt szoftver felhasználásával működik.

A KDK kettős céllal jött létre. Mint repozitórium, a TK-ban létrejött kutatási nyersanyagok digitális másolatait archiválja és kutatási adatokat szolgáltat, elsősorban TK-s kutatóknak. Mint sokrétű és sokfajta metaadatot szervező rendszer pedig kutatásokról szolgáltat alapinformációkat; pl. a kutatás témájáról, résztvevőiről, idejéről, finanszírozójáról stb. lehet tájékozódni.

Az MTA TK négy intézetében (Jogtudományi, Kisebbségtudományi, Politológiai, Szociológiai) létrejövő nyersanyagok, dokumentumok, adatok nagyon sokrétűek. Találhatunk köztük táblázatos adatbázisokat (kérdőíves felmérésből, sajtószemléből, stb.), interjúszövegeket, fókuszcsoporthoz tartozó kutatásokból származó videófelveteleket, fényképeket, terepnaplót, stb. Olyan szoftverre volt szükségünk, amely tehát dokumentumszinten tud metaadatokat kezelni és kereshetőséget biztosítani, másfelől a dokumentumokat gyűjteményekbe tudja rendezni, és egységesen megjeleníteni. Szükségünk volt egy differenciált hozzáférési beállításokat kezelni tudó szoftverre. Alapvetően egy olyan digitális archívumot terveztünk, amely a kutatóknak közvetlenül letölthető, jó minőségű kutatási anyagokat kínál.

Ahogy már fenn bemutattuk, a társadalomtudományos mező Európában és világszerte dinamikusan fejlődik és mindig változó képet mutat. Tehát nem létezik (még?) Európában egy olyan társadalomtudományos archívumi tér, ideértve a kanonizált és elfogadott digitális megoldásokat is, amelyben könnyen és magától értetődően el tudná magát helyezni és el tudna kezdeni működni egy új, komplex társadalomtudományos adatokat archiválni kívánó intézményi archívum Magyarországon.

A hasonló intézmények európa- és világszerte általában saját igényeikre szabott informatikai megoldásokat használnak digitális repozitóriumukban. Létezik már 1-2 olyan kész, angol nyelvű megoldás is, amely sokfajta felmerülő igényt ki tud szolgálni (Dataverse⁹, CKAN¹⁰).

Az MTA tudományos közleményeket tároló repozitórium, a REAL, szintén a nyílt forráskódú Eprints szoftverrel működik. A viszonylag könnyű kezelése, a magyarra már lefordított felülete, és a jó konfigurálhatósága miatt, valamint az MTA repozitóriumok közötti átjárhatóság biztosítása érdekében 2013 elején úgy döntöttünk, hogy Eprints alapokon fejlesztjük a KDK repozitóriumot.

A repozitórium kialakítása 2013 őszén indult el a SZTAKI segítségével. A tudományos publikációkra tervezett Eprints-en alapvető változásokat eszközöltünk. A legfontosabb a kutatási gyűjtemény, mint egység bevezetése volt, melyhez az Eprints ReCollect bővítményét használtuk fel. A ReCollect a UK Data Archive és a University of Essex fejlesztése, és a következő főbb változtatásokat adja a rendszerhez: a kutatási gyűjtemény típushoz egy új, egyedi nézetet rendel, amely sok kapcsolt fájl esetén is jól áttekinthető. A gyűjtemény feltöltésére szolgáló workflow-t is jelentősen leegyszerűsíti, miközben hozzáadja azokat metaadat-mezőket, amelyek a DataCite, INSPIRE és DDI alapján a kutatási adatok leírásához elengedhetetlenek (pl. időbeli és geográfiai lefedettség), végül az új leíró mezők szerinti kereshetőséget is beállítja.

⁸ <http://voices.osaarchivum.org>

⁹ <http://datascience.iq.harvard.edu/dataverse>

¹⁰ <https://orbital.blogs.lincoln.ac.uk/2012/09/06/choosing-ckan-for-research-data-management>

A fentiekén túl még számos további metaadatmezőt is létre kellett hozni, hogy a KDK életének első néhány hónapjában begyűjtött kutatásleírások átvehetőek legyenek. Ezek alapján a kutató már keresés és böngészés közben láthatja, hogy az adott anyag milyen és mennyi dokumentumból áll, milyen a lefedettsége, milyen adatgyűjtési módszereket használtak, stb.

További újításként kereszthivatkozási lehetőséget hoztunk létre, mellyel a kapcsolódó tudományos publikációk és a kutatási gyűjtemények mindkét irányból hivatkozhatóak.

A kutatási gyűjtemény permanens URL-lel rendelkezik, ezáltal egy tudományos publikációban az adatokra való hivatkozás lehetségessé válik. A későbbiekben minden gyűjteményt DOI-val szeretnénk ellátni, egyelőre az EPrints saját URL kiosztási módszerét használjuk. A kutatási adat hivatkozása ma még nagyon egyszerű, nem részletekbe menő, és nem megoldott a közvetlen kapcsolat a konkrét adat és az azt elemző szövegrész között; mégis egy olyan lehetőség bevezetése, amellyel ma Magyarországon még alig élnek, de nemzetközileg egyre elfogadottabbá válik. Ennek megoldásával a Research Data Alliance (RDA) külön munkacsoportja¹¹ valamint az egyik legnagyobb európai társadalomtudományos adatarhívum, a UK Data Service egy projektje¹² is foglalkozik jelenleg.

A gyűjtemények egyes dokumentumainak más és más hozzáférési beállításai lehetnek. A metaadatok mindig nyilvánosak (kivéve egy mezőt, amelyben esetleges szenzitív adatokat tárolunk, pl. egy interjúalany nevét). Társadalomtudományos kutatási anyagok számos esetben szenzitív adatokat tartalmaznak, amelyekhez való hozzáférést korlátozni kell. Ezért a KDK-ban többfajta hozzáférési beállításokat dolgoztunk ki, a SZTAKI Repozitóriumhoz hasonlóan. Vannak gyűjteményeink, amelyek bármelyik honlaplátogató előtt nyitottak, vannak csak MTA TK-s felhasználói adatokkal rendelkezők számára letölthetőek, és vannak olyanok, amelyekhez egyedi, kutatói döntés után férhetnek csak hozzá az érdeklődők.

Mivel a KDK repozitóriumában a személyek beazonosításához megfelelő kutatási anyagok csak a kutatásban résztvevők kifejezett beleegyezési engedélyével tárolhatók, az engedéllyel nem rendelkező dokumentumok esetében a szenzitív adatokat leválasztjuk a többi adatról. Ez egy komplex anonimizálási folyamat része, ahol a folyamat minden lépését dokumentáljuk, és a személyes adatokat tartalmazó dokumentumokat zártan tároljuk. Amennyiben a későbbiekben mégis lehetségessé válik az adatok újbóli összerakása, a dokumentáció segítségével megoldható.

Ez az eljárás megfelel a Data Documentation Initiative¹³ -Lifecycle (DDI-L) ajánlásnak. A DDI-L alapján a metaadat dokumentálása nem a kutatás *után*, hanem már közben történik, minden, az adatok milyenségét befolyásoló lépcsőben. Gyakorlatilag tehát már a kutatás létrejöttékor érdemes elkezdni rögzíteni az adatokkal kapcsolatos információkat. A 2. ábra szemlélteti a DDI-L működési sémáját, és látszik, hogy melyek azok a lépések, ahol ezen ajánlás szabályai szerint dokumentálni kell azt, ami befolyásolta az adatok létrehozását és milyenségét.

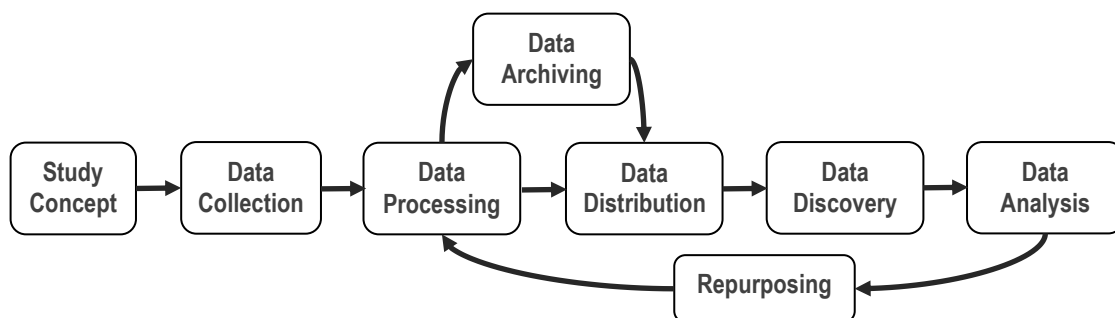
A dokumentáció gépileg is feldolgozható, és számos olyan informatikai fejlesztéssel találkozhatunk, amely a DDI szerint dokumentált metaadatokat kezeli, tárolja, elemzi és jeleníti meg komplex módon. Léteznek olyan szoftverek, amely segítségével a metaadatok és maguk az adatok egy közös felületre rendeződnek ezáltal, ahol látszódik az adatok keletkezésének és kezelésének, változtatásának több eleme.

¹¹ <https://rd-alliance.org/group/data-citation-wg.html>

¹² <http://ukdataservice.ac.uk/about-us/projects/digital-futures/details.aspx>

¹³ <http://www.ddialliance.org>

A DDI és a DDI-L jelenleg leginkább kvantitatív adatok esetében használatos, de már kvalitatív adatok dokumentációjánál is felbukkan.¹⁴ Alkalmazásukkal hatékonyan és nemzetközi összehasonlítást és használhatóságot biztosító módon lehet dokumentálni a kutatási adatok, dokumentumok élettörténetét.



2. ábra. A DDI-L működési sémája¹⁵

A DDI-L bevezetéséhez a kutatók együttműködésére van szükség. A társadalomtudományokban jelenleg sokféle munkamegosztási modell létezik; léteznek olyan projektek, amelyek esetében elengedhetetlen már a kutatás közben dokumentálni az adatok létrejöttét és manipulálását, hogy a kutatótársak is dolgozni tudjanak az adatokkal. Más projektek esetében, főleg egy- vagy néhány személyes kutatások esetében ez ritkábban valósul meg, ezért később már követhetlenné válik számos dokumentum létrejötte és életútja. A kutatási adatmenedzsmentnek mint kutatók felé támasztott eljárásnak a bevezetése¹⁶ egy módja lehet annak, hogy később olyan dokumentumok, adatok álljanak rendelkezésre, amelyeket el lehet helyezni egy repozitóriumban és mások újra tudják elemezni.

4. Összefoglalás

A rendelkezésre álló technológia ma már lehetővé teszi, hogy kevés hardver erőforrással is gyors repozitóriumi megoldásokat hozhassunk létre rövid idő alatt. A probléma inkább ezen repozitóriumok használata, amely technológiai, emberi és szervezeti kihívásokat egyaránt rejt. Még csak kialakulóban van a „best practice” a kutatási adatok kezelésére, és hasonlóképpen kutatások nyilvántartása, a közreműködők, intézmények, támogatók és újrafelhasználók kapcsolatrendszerének nyilvántartása sem általános még. Viszont ezzel egyidejűleg a kutatókat fel kell készíteni ezekre az új kihívásokra, meg kell ismertetni velük az open access jelentőségét, és azt, hogy közleményeiket másképpen érdemes ezután publikálniuk. Az intézményeknek pedig folyamatos tevékenységként be kell vezetniük a kutatási adatok megőrzését, dokumentálását és gondozását, amely újfajta speciális szaktudást igényel. Mindezek teljesülése esetén élvezhetjük csak a tudományos munka új lehetőségeit, arathatjuk le ezek gyümölcseit a támogatások, kapcsolatok vagy újrafelhasználás területein.

A társadalomtudományos adatok, mind a kvantitatívok, mind a kvalitatívok, többlépcsős tisztítási, összevonási adatkezelésen esnek keresztül, mielőtt felhasználják őket a kutatók. A nyers adatokat gyakorlatilag soha nem használják a kutatók, tehát az, amire adatként

¹⁴ <http://ukdataservice.ac.uk/about-us/projects/digital-futures/details.aspx>

¹⁵ forrás: http://odaf.org/papers/DDI_Intro_forNSIs.pdf

¹⁶ Egy jól használható kézikönyv az adatmenedzsmentről: Corti et al (2014) [12].

tekintünk, a kutatói adattisztítási eljárás eredménye. Ez az adatkezelés nem csak másodfelhasználás előtt, hanem már az első felhasználáskor megtörténik. Természet- és társadalomtudományos kutatási adatok és nyersanyagok esetében a kutatók beszámolójára, emlékezetére és (főleg régebbi kutatási anyagok esetében) sokszor levéltári munkára is szükség van, hogy egy-egy kutatás létrejöttének és sorsának aspektusait megismerjük. Az ilyen feltáró munka törvényszerűen csak egy-egy olvasatát adhatja egy kutatás történetének, amelyet a DDI-L keretei között is lehet többé-kevésbé dokumentálni. Az adatok létrejöttét és milyenségét (és nem csak a társadalomtudományosokét) alapvetően meghatározza pl. a felhasznált módszer és a kutatási kérdéseket motiváló érdeklődés; ezeket mind befolyásolják többek között a személyes motivációk, az intézményi prioritások, a tudományos hagyományok, valamint a tudományos mező specifikus működése egy adott időben és tudományterületen [9][10]. Minden tudományos diszciplínában a kutatási adatok és a valóság közötti viszonyt a kutatók munkájuk (elemzéseik, írásaik) során újra és újra létrehozzák [11].

Az idők során általában változik ez a folyamat, és a kutatási adatok más és más valóságot írnak le, más korokban a kutatók más következtetéseket vonnak le, másra helyezik a hangsúlyt, mást tartanak irreleváns és nem elemzendő adatnak. Maguk az adatok is átalakulnak, például ugyanazt az adatbázist felhasználva a társadalomtudósok új változókat hoznak létre a korábbi adatokból, régi interjúk alapján új tudást termelnek, stb. Ahhoz, hogy a kutatói, az értelmezési szabadságot minél inkább segítsük, digitális repozitórium tervezőiként és építőiként arra kell törekedjünk, hogy egy kutatás minél több szeletét minél több nézőpontból mutathassuk be. A szűkös keretek között jelenleg most csupán arra van lehetőségünk az MTA TK Kutatási Dokumentációs Központjában, hogy rugalmas és széles körű metaadatolást tegyünk lehetővé és lehetőleg minél több dokumentumot szerezzünk be a kutatóktól, továbbá az adatokhoz kapcsoljuk az azokat értelmező tanulmányokat. De a jövőben olyan digitális platformokat is el lehet képzelni, ahol a tudósok együtt gondolkodva értelmezik az adatokat, ahol az eredmények és az adatok szimbiózisban mutatkoznak, és ahol ezáltal a tudományos adatokat, történetüket, létrejöttüket, és elemzésük sokszínűségét és sokrétűségét is be lehet mutatni.

Irodalomjegyzék

- [1] Mauthner, N., Parry, O. and Backett-Milburn, K.: The data are out there, or are they? Implications for archiving and revisiting qualitative data', *Sociology*, 32/4 (1998), 733-45.
- [2] ORCID, <http://orcid.org/>
- [3] DataCite, <https://www.datacite.org/>
- [4] MTA Open Access rendelkezés, <http://real.mtak.hu/eprints/mandate.html>
- [5] The Open Archives Initiative Protocol for Metadata Harvesting, 2008-12-07, <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>
- [6] ROAR: Registry of Open Access Repositories, <http://roar.eprints.org>
- [7] The Directory of Open Access Repositories – OpenDOAR, <http://www.openoar.org/>
- [8] OpenAIRE, <https://www.openaire.eu/>

- [9] Bourdieu, Pierre: The Peculiar History of Scientific Reason. *Sociological Forum*. Volume 6, Issue 1, March. (1991), 3–26.
- [10] Kuhn, Thomas: *The Structure of Scientific Revolution*. Chicago: University of Chicago Press. (1962).
- [11] Latour, Bruno - Woolgar, Steve: *Laboratory Life. The Construction of Scientific Facts*. Princeton University Press. (1996).
- [12] Corti, Louise - Van den Eynden, Veerle - Bishop, Libby - Woollard, Matthew: *Managing and Sharing Research Data. A Guide to Good Practice*. Sage. (2014).