METSZETEK
Vol. 10 (2021) No. 3
ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2
www. metszetek.unideb.hu

TEMATIKUS TANULMÁNYOK – Digitális szociológia

Új utak a társadalom megismerésében[1]

A donáció alapú digitális adatgyűjtésben rejlő lehetőségek

KMETTY ZOLTÁN[2]

ABSZTRAKT

Napról napra egyre több digitális adat keletkezik és egyre több társadalomtudományi elemzés használ Twitter, Instagram vagy akár Facebook-adatokat. A „big data" jelenség kapcsán felmerülő társadalomtudományi lehetőségeket és dilemmákat számos nemzetközi és hazai tanulmány végigjárta már – de az „adathoz jutás" kérdésével csak érintőlegesen foglalkoztak ezek a tanulmányok. Az adathoz jutás pedig egyre nehezebbé válik. Mit tehetünk abban az esetben, ha a piaci szereplők lezárják a platformjaikat és ha mégis találunk elérhető adatot, akkor a kutatás-etikai tanács parancsol nekünk megálljt? A válasz egyszerű: forduljunk a felhasználókhoz és tőlük kérjük el az adatokat. Ezt a megközelítést nevezi a szakirodalom adatdonációnak. A tanulmányban részletesen bemutatjuk az adatdonációs megközelítést külön kitérve arra, hogy a jelenlegi nagy nyugati platformok esetében milyen adatokhoz férhetnek hozzá a kutatók a felhasználókon keresztül. Az adatdonációs hozzáférés gyakorlati megvalósíthatóságát egy hazai pilot kutatás alapján mutatjuk be.

KULCSSZAVAK: adatgyűjtés, adatdonáció, közösség média, big data, Facebook

ABSTRACT

New ways in exporting Society
The potential of donation.based digital data collection

More and more digital data is being generated every day, and more and more social science analyses are using Twitter, Instagram, or Facebook data. Many international and national studies have already explored the social science opportunities and dilemmas raised by the phenomenon of ‚big data' - but the issue of ‚access to data' has only been touched upon tangentially. And access to data is becoming increasingly difficult. What can we do if market players close the access to their data, and, if we find data available, the Research Ethics Board tells us to stop? The answer is simple: go to the users and ask them for the data. This approach is what the literature calls data donation. This paper will describe the data donation approach in detail, focusing on how researchers can access data through users on the current major Western platforms. The practical feasibility of data donation access will be illustrated based on a domestic pilot study.

KEYWORDS: data collection, data donation, social media, big data, Facebook

[1] A kutatás az NKFI-től nyert támogatást a Fiatal Kutató Témapályázaton. A kutatás azonosítója: FK128981.

[2] Eötvös Loránd Tudományegyetem, Társadalomtudományi Kar, Szociológia Intézet; Társadalomtudományi Kutatóközpont, CSS-RECENS kutatócsoport, e-mail: kmetty.zoltan@tatk.elte.hu

METSZETEK
Vol. 10 (2021) No. 3
ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2
www. metszetek.unideb.hu

Bevezetés

A kérdőíves survey kutatások uralták a kvantitatív társadalomtudományi kutatások elmúlt 50-70 évét. A kutatók mindig tudatában voltak ennek a módszernek a gyengeségeivel, de mint a rendelkezésre álló legjobb technikát, semmi sem törte meg a hegemóniáját. A közelmúltbeli változások azonban megkérdőjelezték a survey kutatások vezető szerepét. E változások egyik része az egyre nehezebbé váló terepmunka, valamint a csökkenő válaszadási arány; a változások másik motorja pedig új típusú digitális adatforrások megjelenése. A digitális adatok részét képezik azok a nem szándékolt adatok egyrészről amit a különböző általunk használt eszközök felvesznek rólunk, vagy lakásunkról (például: mobiltelefonos adatok helylokációja). A digitális adatok egy másik része a felhasználók által megosztott tartalmak, például tweetek vagy bejegyzések az általuk kedvelt helyekről, vagy más, közösségi médiában megjelenő reakciók, interakciók. Társadalomtudományi szempontból a közösségi média adja a digitális adatok egyik legérdekesebb típusát. A nagy üzleti cégek számára a „Social Listening" része a mindennapos üzleti folyamatnak, mivel a felhasználók igényeinek gyors megválaszolása vagy a fogyasztói vélemények megértése elengedhetetlen részei a fogyasztói elkötelezettség mérésének. Az „ipar" igényei implikálják a közösségi média adatok feldolgozásának gyors fejlődését, de ez az igény a kutatói oldalról is jelentkezik. Egyre több társadalomtudományi elemzés használ Twitter, Instagram vagy akár Facebook-adatokat.

A „big data" jelenség kapcsán felmerülő társadalomtudományi lehetőségeket és dilemmákat számos nemzetközi és hazai tanulmány végigjárta már (Lazer – Radford 2017, Csepeli 2015, Dessewffy – Láng 2015, Kmetty 2018, Ságvári 2017) – de az „adathoz jutás" kérdésével csak érintőlegesen foglalkoztak ezek a tanulmányok. Ez részben azzal magyarázható, hogy látszólagosan rengetek digitális adat érhető el szinte bárki számára – mindenhonnan ömlenek ránk az adatok. De ha jobban körbejárjuk a kérdést, akkor egyértelművé válik, hogy a helyzet korántsem egyszerű. Egyrészről egyre több (piaci) szereplő ismeri fel a saját adatvagyonának értékességét és korlátozza a külső szereplők adat hozzáférését. Másrészről a(z Európában) szigorodó adatvédelmi szabályozások miatt is szűkül a kutatási célból elérhető digitális adatok köre. Mit tehetünk abban az esetben, ha a piaci szereplők lezárják a platformjaikat és ha mégis találunk elérhető adatot, akkor a kutatás-etikai tanács parancsol nekünk megálljt? A válasz egyszerű: forduljunk a felhasználókhoz és tőlük kérjük el az adatokat (Halavais 2019). Ezt a megközelítést nevezi a szakirodalom adatdonációnak. Ebben a megközelítésben követlenül fordulunk a felhasználókhoz – kihagyjuk a piaci cégeket (platform szolgáltatókat) és a felhasználók beleegyezésével kezdjük el használni az adataikat. Ezzel a megoldással nemcsak az adatfelvételünknek teremtünk egy nagyon tiszta jogi környeztet, hanem azt is lehetővé tesszük, hogy akár további kérdőíves adatokat is gyűjtsünk a kutatásba bevont személyektől.

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

www. metszetek.unideb.hu

A tanulmány első felében bemutatjuk a digitális adathozzáférés jellemző csatornáit a megközelítések előnyeivel és hátrányaival. Ezt követően részletesen bemutatjuk az adatdonációs megközelítést külön kitérve arra, hogy a jelenlegi nagy nyugati platformok esetében milyen adatokhoz férhetnek hozzá a kutatók a felhasználókon keresztül. A tanulmány utolsó részében pedig bemutatunk egy olyan egyedülálló hazai pilot kutatást, amely adatdonációs stratégiát használt az adatgyűjtés során.

Digitális adathozzáférés

JELLEMZŐ UTAK

A digitális adatokhoz való hozzáférés jellemző módját az API-k (Application Programming Interface) jelentették. Az API-k alapvetően a számítógépek egymás közötti kommunikációját/adatcseréjét segítik azzal, hogy autentikált csatornákon kaput nyitnak egy adatbázis bizonyos részéhez. Ha például szeretnénk feltűntetni a honlapunkon az éttermünk „google review" értékelését, akkor a honlapunkba tudunk egy olyan kódot integrálni, amely bizonyos időközönként lekéri az aktuális értékeléseket a Google erre dedikált API-án keresztül. Ezt a hozzáférést azonban nemcsak gépek közötti kommunikációra lehet használni, hanem adatok kutatási célú kinyerésére is. A nyilvános API-k egyszerű hozzáférést biztosítanak nagy mennyiségű adathoz, de az adatok minősége változó és az is változó, hogy melyik platform menynyi adatot ad az API-kon keresztül. Bár bizonyos platformok esetekben – mint például a Twitter – ez a hozzáférési mód továbbra is az egyik leghatékonyabb adatelérési út, de más közösségi oldalak, például a Facebook vagy az Instagram esetében ezt a hozzáférési módot leállították vagy drasztikusan megnehezítették a platformok tulajdonosai (Breuer et al. 2021). Az API-k lezárása elsősorban a Cambridge Analytica botrány következménye, de a szigorodó adatvédelmi környezet mindentől függetlenül is abba az irányba terelte a platformokat, hogy szűkítsék az adathozzáférést. Ennek kapcsán Freelon (2018) azt írta, hogy a „Computational Social Science" belépett a „Post-API" korba, Bruns (2019) pedig ezt az egész helyzetet „APIcalypse"-nek nevezte. Mások, mint Tromble (2021) vagy Puschmann (2019) viszont kiemelik ennek a folyamatnak a pozitív hatását, miszerint végre véget ért ezzel a „közösségi média kutatásának vadnyugata".

A nehezedő adathozzáférési környezetben új modelleket kellett kidolgozni a digitális adatokhoz való hozzáférés érdekében.

A NetGain Partnership (Shapiro et al. 2021) által publikált tanulmány két nagy ágát különbözteti meg a digitális adathozzáférésnek: azok a megközelítések, amelyek együttműködnek a platformokkal és azok, amelyek nem.

A platform együttműködés a következő modellekre jellemző (Shapiro et al. 2021: 28):

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

1. **API:** A Cambridge Analytica botrányig ez a hozzáférési mód volt a legelterjedtebb. Bár nagy mennyiségű adathoz lehetett egyszerű módon hozzáférni az API-kon keresztül, de a megközelítésnek voltak hátrányai is: általában kevés adatunk volt a megfigyelt emberekről és arról is, hogy azok az adatok, amihez hozzáférünk hogyan viszonyulnak a teljes adatmennyiséghez. Szintén gondot jelent az API-alapú hozzáférés esetében, hogy a megváltozó adatgeneráló algoritmusok jelentősen gyengíthetik a kapott eredmények érvényességét.
2. **Differenciált adatvédelem (Differential Privacy – DP):** A differenciált adatvédelemnek több megközelítése van. A digitális adatok kapcsán azokat az adathozzáférési módokat értik ez alatt, amikor az adatokra valamilyen zajt tesznek, vagy feltöltik az adatokat véletlenszerű válaszokkal vagy válaszolókkal. Ez valamelyest rontja az adatminőséget és megnehezíti azokat az elemzéseket, amelyekben alapvetően gyenge összefüggés van a változók között. Tipikus példája ennek a megközelítésnek a „Social Science One”[3] project, amelynek keretében kutatók elfogadott kutatási tervekkel hozzáférhetnek zajjal terhelt Facebook-adatokhoz (King – Persily 2020).
3. **Platformok direkt adatmegosztása publikálási megkötésekkel:** Ez egy viszonylag ritka adathozzáférési mód, nagyon szoros kapcsolatot feltételez a kutató és a platform között. Az ilyen típusú hozzáférésből publikált eredmények nem reprodukálhatók és etikailag erősen megkérdőjelezhetők.
4. **Kontrollált környezetben zajló hozzáférés**: Ezt a hozzáférési mód elsősorban az intézményi szereplőkre (pl. kormányzatok, statisztikai hivatalok) jellemző. Az adatszolgáltató kialakít egy olyan kontrollált hozzáférési környezetet (pl. adatszoba), amiben ellenőrízheti, hogy a kutatók milyen adatokat dolgoznak fel és ezekből milyen elemzéseket készítenek. Ha a kutatási eredmények adatvédelmi szempontból nem aggályosak, akkor engedheti az adatszolgáltató azok közlését. Ez az adatgyűjtési megközelítés nem jellemző a közösségi média platformok esetében.

A platformtól független adatgyűjtési módok a következők (Shapiro et al. 2021: 28):

1. **Web-scrapelés:** A web-scrapelés egy nagyon jellemző adatszerzési megoldás a digitális adatok esetében. A web-scrapelés gyakorlatilag a honlapokon és online oldalakon fent lévő információk nyers letöltését és adatbázisba rendezését jelenti. A közösségi média adatok scrapelése kapcsán megfogalmazódnak technikai és jogi dilemmák is (Mancosu – Vegetti 2020, Boeschoten et al. 2020). A közösségi oldalak egyrészről rendszeresen változtatják az oldal kinézetét, szerkezetét, ezért az oldalak leszedését végző kódokat folyamatosan karban kell tartani. Ez a konzisztens adatgyűjtést nagyon költségessé tudja

[3] https://socialscience.one

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

tenni. Másrészről a platform tulajdonosok különböző jogi eljárásokat indíthatnak a kutatókkal szemben. Amerikában a számítógépes visszaélésekkel (CFAA) kapcsolatos törvény nyújt erre lehetőséget, Európában a GDPR szabályozás. A platform tulajdonosok arra hivatkozhatnak ezekben az eljárásokban, hogy a kutatók megsértik a felhasználó feltételeket (Terms of Service – TOS). A scrapelés történhet a felhasználók engedélye nélkül és engedélyével. Utóbbira jó példa a GESIS kutatása, amelyben a kutatásban résztvevők feltelepítettek a gépükre egy böngésző kiegészítőt, ami folyamatosan scrapelte a Facebook-falukat, amikor használták az oldalt (Haim – Nienierza 2019). Ez a megoldás GDPR szempontból maximálisan megfelelő, de ettől függetlenül szembe mehet a platform felhasználási feltételeivel.

2. **App-scrapelés:** Az appok scrapelése még bonyolultabb, mivel az appok általában egy zárt környezetben futnak, amiben nehéz hozzáférni az adatokhoz. Ehhez speciális programokat kell telepíteni a felhasználóknak a gépére/telefonjára, az adatgyűjtéshez. Ez nagyon szoros együttműködést feltételez a felhasználóval.
3. **Adatdonáció**: Ebben a modellben a kutatók megkérik a résztvevőket, hogy osszák meg velük a platformon tárolt adatokat. A nagy platformoknak a GDPR törvény megfelelés miatt lehetőséget kell adni a felhasználóiknak, hogy elérjék és letöltsék a róluk tárolt adatokat, adatcsomagok keretében (data download packages – DDP). A nagy nyugati platformok, mint a Google, a Facebook, az Instagram a WhatsApp vagy a Netflix felhasználóbarát módon adnak lehetőséget az adataink elérésére és letöltésére. Ezeket az adatokat amellett, hogy letöltheti a felhasználó akár további is oszthatja. Ez lehetőséget teremt a kutatóknak, hogy teljesen tiszta jogi környezetben férjenek hozzá közösségi média adatokhoz.

ADATDONÁCIÓ

A vállalatok helyett a felhasználókkal való együttműködés fő előnye, hogy ez átláthatóbbá teszi az adatgyűjtési folyamatot és az adatok felhasználását azok számára, akiknek az adatait használják. Ebben a kutatási megközelítésben nem okoz problémát a felhasználóktól beleegyező nyilatkozatokat gyűjteni. Pontosan elmagyarázhatjuk nekik, miről szól a kutatásunk, mire fogjuk használni az adataikat, hogyan tároljuk majd az adatokat és hogyan biztosítjuk az anonimitás. Az adatgyűjtés során láthatják és akár kontrolálhatják azt, hogy milyen adatokat osztanak meg a kutatókkal (Breuer et al. 2021).

Az adatdonációs megközelítés mivel szoros együttműködést igényel a résztvevőkkel azt is lehetővé teszi, hogy a digitális adatgyűjtés mellett „klasszikus" kérdőíves adatokat is gyűjtsünk a felhasználóktól. A két adatgyűjtési mód ötvözése kuta-

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

tói szempontból egy nagyon értékes adatbázist eredményezhet (Stier et al. 2020). A survey kutatások jellemző problémája, hogy a társadalmi elvárások miatt a kérdezettek gyakran torzítottan válaszolnak a kérdőíves kutatás kérdéseire, de akár felmerülhetnek visszaemlékezési problémák is bizonyos szituációkban (Scharkow 2016). A digitális adatok esetében bár szintén felmerülhetnek bizonyos torzítások, mint például az algoritmusok szerepe – mivel beavatkozás mentesek, nem befolyásolják őket a kutatói percepciók. Ezzel szemben hátránya lehet a digitális adatoknak, hogy sokszor csak a szereplők cselekedeteit látjuk, nem tudunk semmit róluk – hiányoznak a demográfia információk. Ezek a demográfia adatok viszont pótolhatók a survey kutatásból így át tudjuk hidalni a digitális adatoknak ezt a gyengeségét. A két adatforrás kombinálása pedig olyan összefüggéseket is fel tud tárni, ami külön-külön nehéz, vagy akár lehetetlen is lenne.

Az adatdonáció további előnye az, hogy a kutatókat nem kötik az API-k. Az API-k bár egyszerű adatgyűjtést biztosítanak más szempontból korlátozzák az elérhető adatok körét. Csak azokat az információkat érjük el az API-kon keresztük, amit a platform szolgáltatók fontosnak gondolnak. Mivel a legtöbb API nem kutató céllal készült ezért sokszor kapunk számunkra irreleváns adatokat és az igazán releváns adatok pedig hiányoznak. Az adat donációs modellben viszont látunk mindent, amit a felhasználó engedélyez. Az API-alapú közösségi média adatgyűjtések ráadásul általában csak a nem privát tartalmakhoz adnak hozzáférést, ezzel szemben az adat donációs modellben hozzáférhetünk a privát tartalmakhoz is.

További előny az is, hogy akár kis mintanagyság mellett is, az ezzel a módszerrel gyűjtött adatok meglehetősen részletesek és gazdagok lehetnek. Képesek lehetünk akár egy felhasználó teljes közösségi média „életútját" visszakövetni. Ez lehetővé teszi a felhasználók viselkedésének részletes időbeli elemzését akár az egyén és akár a csoport szintjén. Megvizsgálhatjuk például a személyes életesemények (például az egyetem megkezdésének) hatását, vagy vizsgálhatunk jelentős helyi/országos/világ szintű történések online lenyomatait.

A digitális adatokkal szemben gyakori kritika, hogy nehezen általánosíthatók az eredmények. Ennek elsődlegesen az az oka, hogy ritkán tudjuk azt, hogy az általunk gyűjtött adatok mennyiben reprezentálják a vizsgálni kívánt sokaságot. Az adatdonációs megközelítés viszont lehetővé teszi, hogy pontos képet kapjunk a mintánk szerkezetéről és ezáltal az is lehetővé válik, hogy akár a teljes sokaságra is kiterjeszthető eredményeket tudjunk publikálni.

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

www. metszetek.unideb.hu

1. tábla. Az adatdonációs megközelítés előnyei és hátrányai

Előnyök	*Hátrányok*
Részletes adatok a felhasználókról Időbeliség Akár privát megosztások Kombinálhatóság kérdőíves adatokkal Jogilag tiszta környezet Nincs leszűkítve a hozzáférés, mint az API-ban Kontrolálhatjuk, ki kerül be a mintánkba	Aktív közreműködés szükséges a résztvevőktől Kisebb adatbázisok készíthetők, mint egy API-alapú kutatásban Bonyolult az adatokat anonimizálni A kontextust általában nem látjuk, amiben az adat keletkezik A platform szolgáltató megváltoztathatja időnkén, hogy milyen adatok érhetőek el a DDP-ben.

Forrás: Saját szerkesztés: Breuer et al. 2021: 18

Az előnyök mellett fontos megemlítenünk a módszer kapcsán felmerülő limitációkat is. Az API-alapú adatgyűjtés gyors és egyszerű. Az adat donációs adatgyűjtést ezzel szemben alaposan meg kell szervezni. Meg kell győzni a potenciális mintatagokat, hogy vegyenek részt a kutatásban. A részvevőknek pedig aktívan közre kell működni az adatfelvételben, segíteni kell a kutatóknak az adathozzáférésben, ami nem minden felhasználónak triviálisan egyszerű. Az adatgyűjtés nem csak a felhasználókat állíthatja kihívások elé, hanem a kutatókat is. Leginkább az anonimizálás igényel extra odafigyelést, de maga az adatszerkezet bonyolultsága is kihívás elé állíthatja a kutatókat. Magas szintű programozói tudás szükség ahhoz, hogy az egyedi felhasználóktól érkező adatokat össze tudjuk rendezni egy (vagy több) elemezhető adatbázisba. Ez még inkább igaz akkor, ha az adatfelvételi időszakon belül változtat a platform szolgálató az adathozzáférési protokollján. Ez további külön adatintegrációs lépéseket indukál a kutatók részéről.

A közösségi média adatok bár nagyon részletesek, de a kontextusuk sokszor hiányzik. Azt például láthatjuk, hogy kedvel valamilyen tartalmat a felhasználó, de a konkrét tartalmat már nem láthatjuk, mert az az adat már a másik felhasználó „tulajdona" ezért a résztvevő adatletöltési csomagjának ez már nem része. A megfigyelés teljeskörűségének hiánya abban is jelentkezik, hogy 'csak' azokat a tevékenységeket láthatjuk, amikor a felhasználó aktívan tesz valamit, amikor csak „böngész" és „megfigyel" eltűnik a kutatói radarról. Ez például fontos probléma lehet olyan kutatásoknál, amelyek a felhasználók hírkitettségét vizsgálják. Ilyen esetekben lehet érdekes a kutatási dizájn kombinálása web böngésző alapú adatgyűjtéssel is[4].

A limitációk ellenére a DDP-n keresztüli adatgyűjtés az egyik legígéretesebb új módszer a közösségi média tevékenység vizsgálatára. A következő részben rövid át-

[4] Ilyen kombinált adatgyűjtés tudomásunk szerint még nem készült soha.

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

www. metszetek.unideb.hu

tekintést adunk arról, hogy milyen típusú adatokhoz férhetünk hozzá egy adat donációs projektben a Facebook az Instagram és a Google esetében.[5]

FACEBOOK

A Facebook jelenleg a legnagyobb közösségi média oldal és a növekedése töretlen volt az elmúlt 10 évben. 2020 végére 2,8 milliárd aktív használója volt a Facebooknak havonta, és 1,8 milliárd naponta.[6] Minden percben több mint 500 ezer komment születik és 300 ezer státusz frissítés.[7] Magyarországon több mint 5,2 millió Facebook-használó van, ami azt jelenti, hogy az aktív internetezők közel 90% aktív a közösségi oldalon, vagy legalábbis rendelkezik profillal.[8] A magyar felhasználók elsősorban a barátokkal, ismerősökkel való kapcsolattartásra használják az oldalt, de több mint a felhasználók fele a híreket is itt fogyasztja, ami jól mutatja az oldal jelentőségét.

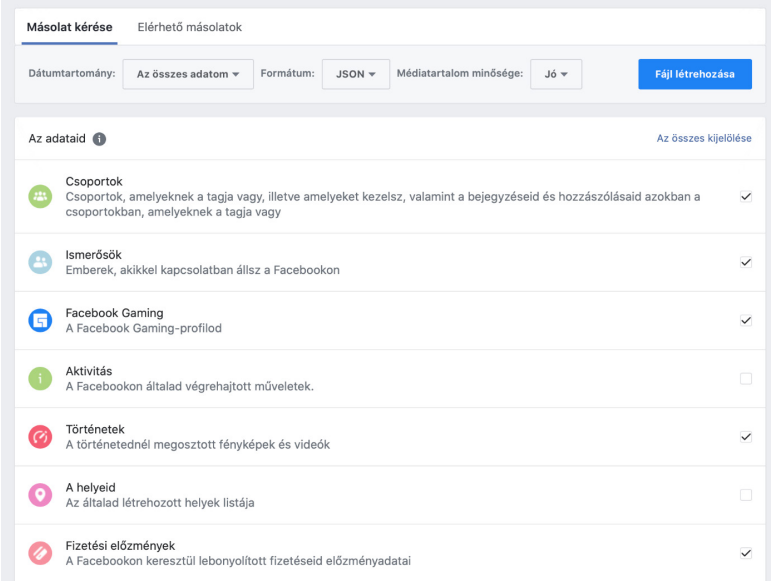
A saját Facebook-adatok a beállítások menüben érhető el, és ezen belül „A Facebook-adataid" almenüre kell kattintani. Ha valaki csak böngészni szeretné a Facebook-adatait akkor a „Hozzáférés az adataidhoz" opciót lehet választani, az adatok letöltéséhez viszont a „Saját információ letöltése" linkre kell kattintani. Itt hozzáférhetünk minden rólunk tárolt adathoz. Az adatkörök folyamatosan bővülnek, ahogy a Facebookon új funkciók kerülnek kialakításra. Elérhetőek olyan alap információk, mint a barátlista, reakciók, kommentek, posztok, vagy a csoporttagság. De itt vannak a profil adataink is, illetve olyan információk például, hogy a Facebook milyen érdeklődési kategóriába sorol minket. A letöltésnél kiválaszthatjuk, hogy milyen tartalmakat szeretnénk letölteni – nem kell a teljes információmezőt letölteni. Három további opció van a letöltésnél. Kiválaszthatjuk az időszakot, a multimédiás tartalom minőségét és a letöltési formátumot. Utóbbi esetében kettő közül választhatunk: html vagy JSON. Utóbbi javasolt inkább, a JSON fájlokat nagyon hatékonyan tudja olvasni a Python és az R is. Az R és Python utalás arra is rávilágít, hogy ahhoz, hogy elemzési formátumra hozhassuk ezeket az adatokat kell a projekt csapatba legalább egy nagyon jól programozó kolléga.

A tanulmány harmadik felében bemutatunk majd egy olyan projektet, amelyben Facebookról gyűjtöttünk adatokat. A Facebook-adatokra épülő adatgyűjtés további technikai részeket ezért itt most nem bontjuk ki, a későbbi részekben fogjuk ezeket ismertetni.

[5] Olyan platformokat mutatunk be, amelyeket a magyarok rendszeresen használnak. A kör bővíthető lenne más platformokkal, mint a Twitter vagy a WhatsApp. Ezeknek azonban alacsony a hazai bázisa.
[6] https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/
[7] https://zephoria.com/top-15-valuable-facebook-statistics/
[8] http://nmhh.hu/dokumentum/187704/lakossagi_internethasznalat_2016.pdf

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

www. metszetek.unideb.hu



1. ábra. A Facebook adatletöltés ablakából egy részlet
Forrás: www.facebook.com

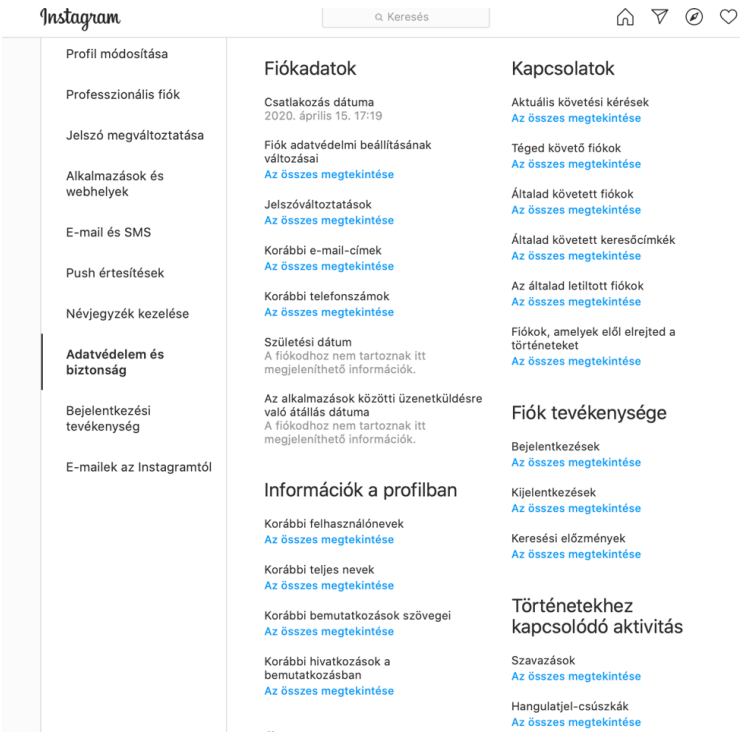
INSTAGRAM

Az Instagram a második legnagyobb közösségi oldal Magyarországon. Az Instagram penetrációja 36% körüli az internetező lakosságon belül.[9] De a fiatalabbak esetében ez az arány még magasabb. Az Instagramon a Facebookhoz képest kevesebb funkció érhető el, több az audiovizuális elem és kevesebb a szöveges adat. Bár az Instagram nem olyan népszerű adatforrás, mint a Twitter a társadalomtudományokban, de bőven vannak olyan tanulmányok, amelyek az Instagram adatokat használják társadalomtudományi elemzésekben (Moreno et al. 2016, Reece – Danforth 2017, Brown et al. 2018, Koltai – Kmetty – Bozsonyi 2021). A fő vizsgált témák a mentális egészséggel, a cyberbullyinggal vagy az önképpel, önreprezentációval kapcsolatosak voltak. Az utóbbi években egyre több tanulmány kezdett el foglalkozni az audiovizuális elemek elemzésével is.

[9] https://nmhh.hu/dokumentum/212534/internet_2019_tanulmany.pdf

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

www. metszetek.unideb.hu



2. ábra. Az Instagram adatletöltés ablakából egy részlet
Forrás: www.instagram.com

A legtöbb tanulmány az Instagram API-t használta az adatgyűjtéshez, de ez az adathozzáférés gyakorlatilag megszűnt, miután a Facebook felvásárolta az Instagramot. Tudomásunk szerint az adatdonációs megközelítést eddig egyetlen társadalomtudományi kutatásban sem alkalmazták. Boeschoten és szerzőtársai (2020) tanulmányukban egy fiktív Instagram-alapú kutatással illusztrálják az adatdonációs megközelítés előnyeit.

A Facebookhoz hasonlóan az Instagram is lehetővé teszi, hogy a felhasználók – a GDPR-rendeletnek megfelelően – hozzáférjenek saját adataikhoz és letöltsék azokat. Az adatvédelmi al-oldalon ellenőrizhetjük, hogy az Instagram milyen típusú adatokat rögzít rólunk. Letölthetők technikai metaadatok (bejelentkezési adatok), profilinformációk, különböző információk a kapcsolatokról, korábbi keresésekről, történetekről vagy megosztott posztokról és képekről. Az időbélyegző és a földrajzi helymeghatározás is sok esetben elérhető.

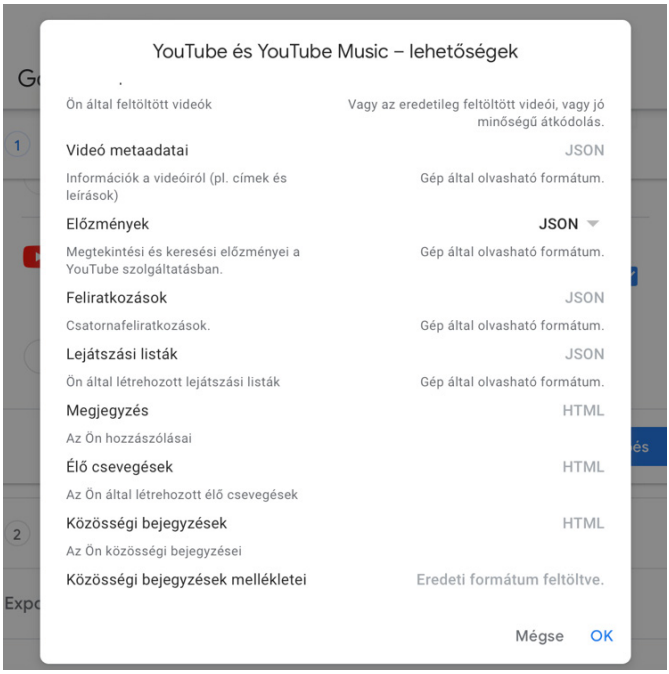
ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

www. metszetek.unideb.hu

Az adatletöltő oldalon csak az adatformátumot tudjuk kiválasztani (html/JSON), de azt nem, hogy milyen adatokat szeretnénk letölteni. Az Instagram a felhasználó teljes adatarchívumát exportálja. Az Instagram által biztosított DDP jelentősen különbözik a Facebook által biztosított DDP-től. Mivel a felhasználó által megosztott összes audiovizuális tartalmat megkapjuk, nagyobb lesz az adatméret, ami megnehezíti bizonyos adatmegosztási sémák alkalmazását.

GOOGLE – YOUTUBE

A Google nem egy önálló platform, hanem inkább egy szolgáltató. A Google szolgáltatásait keresésre, e-mailezésre, adatmegosztásra vagy zenehallgatásra használják. Arról nincs információnk, hogy hányan használják Magyarországon a Gmailt e-mail kliensként, de azt tudjuk, hogy Magyarországon az internetezők több mint 80%-a hallgat zenét vagy néz videókat a YouTube-on.[10]



3. ábra. A Google Takeout adatletöltés ablakából egy részlet
Forrás: www.google.com

[10] https://nmhh.hu/dokumentum/212534/internet_2019_tanulmany.pdf

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

Szolgáltatásain keresztül a Google különböző típusú adatokat tárol a felhasználókról. Egyes szolgáltatásokat csak kevesen használják, így ezek az adatok kutatási szempontból kevésbé érdekesek. Más szolgáltatások nagyon érzékeny adatokat (például e-maileket) tartalmaznak, így ezen adatok gyűjtése nem reális erős lemorzsolódás nélkül. Az egyedi kutatási igények szabhatják meg, hogy milyen adatokra érdemes célozni a Google univerzumban. Ilyenek lehetnek például a következők:

– A felhasználó által kattintott vagy megtekintett hirdetések
– YouTube-aktivitás
– Google keresési előzmények
– Helymeghatározási előzmények.

A Google létrehozta a „Google Takeout" szolgáltatást az adatok letöltésére. Szemben az Instagrammal, itt a felhasználó kiválaszthatja, hogy milyen adatokat kíván letölteni. Az adattípusra kattintás után további opciók válnak elérhetővé. Az adatok formátuma nem egységes, az adattípustól függ. Szöveges adatok esetén a html/JSON a leggyakoribb formátum, de lista típusú adatok esetén a CSV is gyakori. Ha audiovizuális adatokat szeretnénk letölteni akkor is több formátum közül választhatunk.

Egy hazai kísérleti kutatás bemutatása

ADATGYŰJTÉS

A tanulmány eddigi részében bemutattuk, hogy milyen módszerekkel lehet digitális adatkhoz hozzáférni, ezen módszereken belül melyek az előnyei az adatdonációs megközelítésnek és körbejártuk, hogy a három nagy hazai penetrációjú platformon (Facebook, Instagram, Google) milyen adatokhoz és hogyan férünk hozzá. Az adatdonációs megközelítést azonban nem csak elméletben tudjuk bemutatni, hanem egy hazai pilot kutatás tapasztalataival is rendelkezünk. Ez a pilot kutatás 2018 őszén indult az NKFI támogatásával.[11]

A kutatásban azt tűztük ki célul, hogy leteszteljük egyrészről a technikai megvalósíthatóságát egy adatdonációs projektnek, másrészről felmérjük azt, hogy milyen típus társadalomtudományi elemzésekre alkalmas egy így nyert adat. Ebben a fejezetben az előbbi célokat vizsgáljuk, az elemzési lehetőségekre később térünk ki. A projekt első lépésében megvizsgáltuk azt, hogy milyen adatok érhetőek el egyáltalán a Facebookos adatletöltést használva. Ehhez a kiinduló profilokat a kutatócsoport[12] kutatóinak saját profiljai adták. Ez a lépés segített nekünk megérteni, hogy

[11] FK 128981
[12] A kutatócsoportban a következő kutatók vettek részt: Kmetty Zoltán (kutatásvezető), Németh Renáta, Boros Krisztián, Mogyorósi Pálma, Tarnói Csenge, Vancsó Anna, Váry Dániel.

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

egyes tevékenységek hogyan reprezentálódnak a Facebook-adatsorban, mi az, amit láthatunk egy tevékenység kapcsán és mi az, amire nincs rálátásunk. Ezen a ponton döntöttünk arról is, hogy melyek azok az adatok, amelyek mindenképp szeretnénk gyűjteni és melyek azok az adatok, amelyekről le tudunk mondani. Kutató szempontból elsőre jó ötletnek tűnhet minden adatot elkérni. De ez három szempontból is problémás lehet.

- Vannak olyan funkciók, amit kevesen használnak vagy már nem használnak a felhasználók. Az ilyen funkcióra épülő adatokkal nem tudunk mit kezdeni, kvantitatív elemzésben nem használhatók értelmesen. Tipikus példa erre a Facebook-piactér adatsora, vagy a bökés (poke) funkció.
- Vannak olyan adatok, amik nagyon érzékeny információkat tartalmazhatnak. Ilyenek például az üzenetek. Ha olyan tartalmakat kérünk el a felhasználóktól, amit nem szívesen adnak oda, azzal növeljük a kutatás visszautasításának valószínűségét. Mivel a kutatás alapban is nagyon sokat kér a résztvevőktől, fontos találnunk ebben egy működő egyensúlyt. Ezt a logikát követve a privát üzeneteket és a keresési előzményeket nem kértül el végül a résztvevőktől.
- Adatvédelem szempontjából fontos elvárás minden hasonló kutatás kapcsán, hogy maximálisan biztosítsa a résztvevők anonimitását. Az anonimitás biztosítása sokkal nehezebb egy ilyen kutatásban, nem elég a válaszadótól külön fájlban kezelni a nevét és a címét. Az audiovizuális elemek anonimizálása nagyon komplex feladat, amely magában rengeteg erőforrást igényel. Ezért úgy döntöttünk, hogy az audiovizuális tartalmakat nem kérjük el a felhasználóktól. Ez például azzal a pozitív következménnyel is járt, hogy nem kellett extra nagy tárhelyet biztosítani a kutatási adatoknak.

A kutatás lebonyolításával egy professzionális piackutató céget bíztunk meg. A tervezett (és végül meg is valósult) minta nagyság 150 fő volt. A kutatásban egy gyenge életkori kvótát használtunk, annyit kötöttünk ki, hogy a minta legalább 40 százaléka 30 év feletti kell, hogy legyen. A minta nem valószínűségi minta, tehát nem reprezentálja a Facebook-használókat. A célunk a pilot kutatással elsősorban az volt, hogy megvizsgáljuk a kutatás technikai lebonyolításának lehetőségeit és képet kapjunk arról, mire lehet használni a kinyert adatokat. A résztvevőknek rendszeres Facebook-felhasználóknak kellett lenniük (azaz legalább heti rendszerességgel használniuk kellett a platformot). A válaszadók mindegyike magyar volt, az ország keleti részén élt, többnyire nagyvárosban. A terepmunka 2019 áprilisa és szeptembere között zajlott. A kutatási alanyok 3000 forintot kaptak a részvételért. A piackutató cég a saját kutatási adatbázisát használta fel a toborzáshoz, illetve diákszövetkezeteken keresztül is folyt rekrutáció. Azoknak a résztvevőknek, akik beleegyeztek a vizsgálatban való részvételbe, elküldtük a projekt leírását, és kaptak egy meghívót a piackutató cég helyi irodájába.

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

Fontos technikai részlet, hogy a Facebook-adatok, a letöltés elindítása után nem vállnak rögtön elérhetővé. A Facebook alacsony prioritást rendel az adatletöltés mellé, ezért általában lassan készül el az adat. Néha perceket kell várni, de van, amikor órákat is (főleg ha nagy az adattömb). Ezért a kutatási felkérést elfogadókat arra kérte a piackutató cég, hogy a tervezett találkozó előtt egy nappal indítsák el az adatexportálási folyamatot. A folyamat megkönnyítése érdekében a résztvevők részletes leírást kaptak arról, hogyan férhetnek hozzá a Facebook-adataikhoz, és hogyan indíthatják el az exportálást.

Az adatdonációs megközelítés egyik legnagyobb előnyének emeltük ki, a tiszta jogi környezetet és a jól érvényesíthető adatvédelmi szempontokat. Ennek fontos alapja a résztvevők beleegyező nyilatkozata a kutatásba, illetve az adatkezelési elvek elfogadása. Ehhez igazodva a vizsgálat megkezdése előtt a résztvevőknek alá kellett írniuk egy részvételi beleegyező nyilatkozatot.

Ahogy korábban már említettük a kutatási alanyoknak személyesen be kellett menni a piackutató cég egyik irodájába. Elviekben elképzelhető egy olyan kutatási dizájn is, ahol a résztvevők feltöltik a Facebook-adataikat egy dedikált szerverre és nem mennek be az irodába. További lehetőség lehet az is, hogy kérdezőbiztosok az otthonukban keresik fel a résztvevőket és a kérdezőbiztos gépére mentik el az adatokat. Azonban mi általunk alkalmazott kutatási megközelítésben fontos elem volt, hogy nem a nyers JSON fájlokat kaptuk meg a kutatást lebonyolító cégtől, hanem egy elő-feldolgozott és lehetőségek szerint anonimizált adatsort. Ez az előfeldolgozás viszont csak jól előkészített környezetben működött stabilan.

A munka során kifejlesztettünk egy R kódot, ami a letöltött JSON fájlokat előfeldolgozta, átalakította CSV fájlokká és eltávolította az adatokból a megtalált neveket. Két dolog motivált minket, amikor elkezdtük ennek a kódnak a megírását. Egyrészről fontos volt, hogy biztosíthassuk a résztevőknek azt a jogot, hogy belenézhessenek a letöltött adataikba és akár ki is törölhessenek bizonyos tartalmakat. JSON fájlok esetén ez nagyon nehézkesen lett volna megvalósítható, viszont a CSV fájlok egyszerűen olvashatók (akár excelben is). Ezzel a megoldással ezt a lehetőséget biztosítottuk a résztvevőknek. Fontos ennek kapcsán megjegyezni, hogy végül senki nem törölt ki adatok a letöltött archívumából. A másik motiváció az volt, hogy ne kelljen személyes adatot kezelnünk az adatfelvétel után, a személyes adatok kezelése teljes egészében a piackutató cég feladata legyen. Ehhez a kutatási alanyok nevét, és a Facebook-barátaiknak a nevét elfedtük az algoritmussal, hashelés segítségével (lásd hasonló kapcsán: Mancosu – Vegetti 2020). Ez a hashelés sokkal jobb megközelítés, mint az egyszerű törlés, mivel így anonim módon, de nyomon követhetjük egy adott Facebook-baráttal, mikor és milyen interakció történt. Az összes adatot ugyanazzal a hashelő módszerrel fedtük le – ez biztosította azt, hogy ugyanannak a személynek minden adatbázisban ugyanaz legyen a kódja is.

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

Az átalakító és anonimizáló R kód tesztelésekor számos problémát azonosítottunk a karakterkódolásokkal és a különböző programverziókkal kapcsolatban. A hasonló problémák elkerülésének érdekében a piackutató cég egy külön ezt a célt szolgáló laptopot használt a vizsgálathoz. Az irodában a résztvevőket arra kértük, hogy a kijelölt laptopon jelentkezzenek be a Facebook-fiókjukba, majd töltsék le az előkészített Facebook-adatprofil-archívumukat JSON formátumban. Ez választ ad, arra a korábban feltett kérdésünkre, hogy miért volt szükség arra, hogy a kutató cég irodájában megjelenjenek a résztvevők. A letöltött adatok átalakítása és anonimizálása nagyon nehezen lett volna kivitelezhető, ha más dizájnt választunk.

Egy érdekes következménye volt a hashelésnek. A kutatásban résztvevők Facebook-barátairól nagyon korlátozott információk állnak csak rendelkezésre. A résztvevővel zajló interakciókat látjuk, de arra csak egy gyenge becslésünk lehet, hogy milyen korú, milyen iskolai végzettségű, vagy milyen településtípuson lakik a barát. Az emberek neve viszont nagyon árulkodó információ a nemmel kapcsolatban. De a hashelés után ez az információ eltűnik. Ezért még a hashelés előtt, az összes Facebook-barátnak megbecsültük a nemét az elérhető magyarországi keresztnév listák alapján. Ez a megoldás több mint 90%-át besorolta az ismerősöknek. Ott volt gondban az algoritmus, ahol becéző nevek jelentek meg a Facebookon (pl. Kovács Eszti Kovács Eszter helyett), vagy az ismerősnek a családi és keresztnevében keveredtek a fiú és női nevek (pl. Ágoston Sára) vagy nem magyar Facebook-ismerőse is volt valakinek. Ez a példa jól mutatja, hogy már az adatfelvétel tervezése során érdemes átgondolni azt, hogy milyen változókra lehet majd szükségünk a későbbi elemzésekben.

Ahogy a fejezet felvezetésében részletesen foglalkoztunk vele, nem a teljes archívumot kértük el a résztvevőktől, hanem szűkítettük az adatok körét. Többek között nem kértük el a multimédia tartalmakat, a Facebook Messenger-üzeneteket, vagy a keresési előzményeket. A szűkített adatsor ellenére a kutatásban gyűjtött adatok a Facebook-tevékenységek széles körét fedik le, mint a bejegyzések, hozzászólások, kedvelések és reakciók, kedvelt oldalak, barátok, profilinformációk és hirdetésekre vonatkozó adatok. Az adatok tartalmi kibontásával a következő fejezetben foglalkozunk.

A Facebook-adataik megosztása mellett a résztvevőknek egy online kérdőívet is ki kellett tölteniük amíg az irodában tartózkodtak. A kérdőív nagyon változatos témákat fedett le, tartalmazott kérdéseket a politikáról, a médiahasználatról, az önreprezentációról, a mentális egészségről, a szabadidős tevékenységekről és a zenei preferenciákról is. A két adatforrás összekapcsolása érdekében ugyanazt az azonosító kódot használtuk az online kérdőívben és a Facebook-adatok tárolásakor. A nyers Facebook-adatokat az anonimizálási folyamat után törölte a kutató cég. A piackutató cég az anonimizált adatokat csak a vizsgálat vezető kutatójával osztotta meg.

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

GYŰJTÖTT ADATOK KÖRE

A viszonylag kicsi, 150 fős minta ellenére egy nagyon komplex és gazdag adatsor keletkezett a kutatásban. Nem célunk, hogy jelen tanulmányban minden apró részletét kibontsuk a gyűjtött adatoknak, de egy átfogó képet mindenképp szeretnénk adni az adatsorról. A tanulmányban azt a struktúrát követjük, ami mentén a Facebook is rendezi az adatokat.

Barátok: A résztvevők Facebook-barátaira vonatkozó adatok tartalmazzák az összes barátjuk listáját, a barátság kezdetének időbélyegét, a barát hashelt nevét és a barát becsült nemét (lásd korábban). Az elutasított és függőben lévő baráti kérésekről, valamint az eltávolított barátokról is rendelkezünk információkkal. A 150 résztvevőnek összesen 115 955 élő barátkapcsolata volt, ami átlagban 773 barátot jelent. A legmagasabb Facebook-barátszám 2840 volt.

Oldalak: Az adatok tartalmazzák a résztvevők által követett összes oldal listáját, valamint egy időbélyeget, amely jelzi, hogy mikor kezdték el követni az adott oldalt. Az oldalak egy részét (a mintánkban szereplő oldalak mintegy kétharmadát) a Facebook 18 különböző kategóriába sorolta (pl. könyvek vagy zenei oldalak). Összességében a vizsgálatunkban részt vevő 150 résztvevő 83 232 oldalt követett. Az egyedi oldalak száma 52 700. A résztvevők átlagosan 562 oldalt követtek.

*Reakciók***:** Ez az adathalmaz tartalmazza a felhasználók összes reakcióját, időbélyegzővel és a reakció típusával (a leggyakoribb a tetszés). Tartalmazza továbbá a reakció célpontját (azaz, hogy az egy barát vagy egy oldal volt-e). A barátok posztjaira adott reakciók messze gyakoribbak (76,5%), mint az oldalak tartalmára adott reakciók. Fontos, hogy csak a metaadatok szintjén van információnk arról a tartalomról, amelyre a felhasználó reagált. Az adatsor nem azt tartalmazza, hogy pontosan mire reagált a résztvevőnk, hanem csak a tartalom típusára vonatkozó információkat. A tartalom 11 típusba sorolható, plusz egy „egyéb" opció. Adatainkban a legtöbb reakciót a posztokra (45,0%) és a képekre (43,8%) adott reakciók jelentették. Az adatbázisban szereplő reakciók száma összesen 1 802 430.

Egyéb Facebook-tevékenységek: A reakción kívül számos más típusú tevékenységet is végezhetnek az emberek a Facebookon. A leggyakoribbak a hozzászólás és a kommentelés, de a felhasználók fényképeket, videókat is feltölthetnek, linkeket oszthatnak meg, játszhatnak, szavazásokat indíthatnak, eseményeken vehetnek részt stb. Összességében 346 407 tevékenység rekordjai szerepel az adathalmazunkban. A leggyakoribb tevékenység a posztolás. A posztoláshoz rendelkezünk a tevékenység időbélyegzőjével, valamint a poszt tényleges tartalmával, a tevékenységben részt vevő barátok (maszkolt) nevével, a felhasználók által esetlegesen megosztott linkekkel és a poszthoz fűzött kommentek számával (ha van ilyen). Ezen túlmenően az adatok további információkat is tartalmaznak, például annak az eseménynek vagy csoportnak a nevét, amelyre/amelyre vonatkozóan a felhasználó posztolt/feltöltött/megosztott.

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

Érdeklődési kategóriák: A Facebook minden felhasználót kategorizál és ezt a kategorizálást használja a hirdetési optimalizálásra. A besorolás teljesen algoritmikus és nem pontosan ismert, hogy milyen adatokat használ fel. Az elérhető információkból arra lehet következtetni, hogy a besorolásban szerepe, van a saját kedveléseknek (p.l: oldalak), a tevékenységeknek, a használt keresési szavaknak és még a barátok preferenciáinak is (DeVito 2017). De mivel az algoritmus egy fekete doboz, így csak a kategorizálás eredményét tudjuk pontosan megfigyelni. Ebben az adathalmazban nincs időbélyegző, csak a kategórianevek elérhetők felhasználónként. A 150 résztvevő 105 642 érdeklődési kategóriához kapcsolódott (18 689 egyedi kategória). A mintánkban szereplő résztvevők átlagosan 704 kategóriához kapcsolódtak. A kategóriák között vannak általánosabbak (pl. étel) és specifikusabbak is (Pizza).

Az adatok rövid bemutatása után arra térünk ki, hogy milyen elemzéseket tud támogatni egy ilyen jellegű adatsor.

ELEMZÉSI LEHETŐSÉGEK

Az előző fejezet jól illusztrálja, hogy milyen gazdag adatbázis keletkezik egy adatdonációs megoldást használó kutatásban. A Facebook-archívum csak az adatok egyik felét fedik le, ezekhez jönnek még a klasszikus survey adatok is. Utóbbi persze nem kötelező, de a dizájn miatt gyakorlatilag értelmetlen kihagyni.

Módszertani oldalról legalább három lehetséges elemzési irányt felrajzolhatunk. Az egyik irány a Facebook és survey adatok összevetése érvényességi oldalról. Erre jó példa az a tanulmányunk, amelyben a Facebook-tevékenységek alapján azonosítható zenei ízlést vetettük össze a kérdőívben mért zenei preferenciákkal (Kmetty – Németh 2020). Az ilyen jellegű kutatások fontos ismereteket nyújtanak a különböző adatsorok érvényességi korlátairól, de segíthetnek abban is, hogy milyen elvek mentén érdemes a Facebook-tevékenységek alapján operacionalizálni egyes kérdéseket.

Egy másik triviális megközelítés, ha a két adatsorral kiegészítjük egymást. A közösségi média aktivitás és depresszió összefüggését vizsgáló tanulmányunkban (Kmetty – Bozsonyi 2021) hasonló logikát követtünk. A kutatásban a depresszió szintjét egy a survey adatbázisban egy kérdőív blokkal mértük. Ez volt a függő változó a modelljeinkben. A Facebook-adatokból pedig olyan indikátorokat alakítottunk ki, amelyek feltételezésünk szerint összefügghetnek a depresszióval, mint például a posztolás gyakorisága, a posztolás időbeli változása vagy a kutatási résztvevők érdeklődési körei. Minden módszernek megvan a saját limitációja arra vonatkozóan, hogy mit lehet és mit nem lehet vele mérni. Ha jól kombináljuk a módszereket, akkor olyan összefüggésekig tudunk eljutni, ami a korábbi kutatási megközelítésekben elérhetetlen volt.

A harmadik irány, amit kiemelnénk elsősorban a közösségi média adatokat használja és nem nyúl a survey adatokhoz, vagy csak nagyon korlátozott mértékben.

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

www. metszetek.unideb.hu

Ezeknek az elemzéseknek a fő fókusza a közösségi média adatok dinamikai szempontú elemzése. Mivel a legtöbb tevékenységhez időbélyeg társul, ezért az adatsort nem csak statikus modellekben lehet vizsgálni. Ezt a logikát követi például az a munkánk (Kmetty et al. 2020) amelyben a közösségi média aktivitás időbeli dinamikáját elemezzük. A Facebook-használókra jellemző, hogy egy szűk időintervallumon belül nem keverik a tevékenységeket, tehát az aktivitásuk „vonatszerű". Elemzésünkben ezeket a vonatokat azonosítjuk és vizsgáljuk azt, hogy mitől függ az, hogy valaki tevékenységi típust vált.

A három felsorolt példakutatás természetesen nem fedi le az összes lehetséges irányt, de jól illusztrálja, hogy milyen elemzésekben gondolkozhatunk akkor, ha hozzáférünk donáció alapú közösségi média adatokhoz. Ha egy kutatáson belül több adatforrást is megoszt velünk a kutatási alany (pl. Facebook + Instagram), akkor az elemzési lehetőségek még egy további dimenzióval bővülnek. Várhatóan a következő években meg fognak jelenni ilyen komplex adatfelvételek is.

Összegzés

A digitalizáció folyamatosan alakítja a társadalmunkat. Megváltoznak a kapcsolattartási szokásaink, a médiafogyasztásunk, a szabadidő eltöltési szokásaink, vagy akár a munkánk jellege is. Ez a változás nemcsak tartalmi szempontból érdekes a társadalomtudományoknak, hanem módszertani szempontból is. A digitalizáció új típusú adatokat generál, a digitális kutatások infrastrukturális környezetének dinamikus fejlődése pedig új módszertani arzenált biztosít a kutatóknak. A digitális adatok azonban a lehetőségek mellett rengeteg kihívással és kérdéssel is járnak. Hogy tudunk sok adatot feldolgozni, kire lesz „reprezentatív" az elemzésünk, működnek-e a standard módszereink, hogyan tudjuk megtalálni a szociológia szempontból érdekes tartalmat egy nagyon zajos adatban? De talán a legfontosabb kérdés az, hogyan tudunk jó minőségű adathoz jutni? Erre a dilemmára vázoltam fel egy lehetséges megoldást, a donáció alapú adatgyűjtést. Ebben a megközelítésben a kutató nem a platform szolgáltatókkal áll kapcsolatban, hanem közvetlenül a kutatási alanyokkal és tőlük „szerzi" be az adatokat. Ez egyrészről egy jogilag sokkal tisztább adathozzáférési mód, hiszen adott benne a résztvevő beleegyezése. Másrészről a hozzáférhető adatok köre is sokkal gazdagabb, mint például egy standard API-alapú hozzáférésnél. És akár arra is van lehetőségünk, hogy egy survey kutatással kombináljuk ezeket az adatfelvételeket. Cserébe sokkal több energiát kell egy ilyen adatfelvétel megszervezésébe belefektetni és a mintanagyságok is jóval kisebbek lesznek.

A GDPR szabályozás gyakorlatilag kikényszerítette az összes nyugati világban aktív közösségi média platformtól, hogy egyszerű módon tegye elérhetővé a felhasználóknak a róluk gyűjtött adatokat. Ez megteremtette annak is a lehetőségét, hogy kutatóként hozzáférjünk Facebook, Instagram vagy Google adatokhoz, ha meg tud-

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

juk győzni a felhasználókat arról, hogy átadják nekünk ezeket az adatokat. A donáció alapú adatgyűjtések még gyerekcipőben járnak, ezért jelenleg kevés gyakorlati példa van ezekre a kutatásokra. Ritka kivétel az az ELTE Társadalomtudományi Karán zajló pilot kutatás, amelyben 150 aktív Facebook-használótól gyűjtöttünk adatokat. Ennek a kutatásnak a technikai lebonyolítását részletesen bemutattuk jelen tanulmányban.

A donáció alapú adatgyűjtés jövője részben a platform szolgáltatókon múlik. Ha engednek a szigorú API hozzáférési szabályokon akkor kisebb lesz az igény donáció alapú adatgyűjtésre. Bár az adatdonációnak számos előnye van, a gyors és nagy adatmennyiséget generáló API-kal nehéz felvenni a versenyt. Az utóbbi évek tendenciái azonban inkább a hozzáférés további szigorítását vetítik előre, nem az enyhítését. Az adatdonációs modell elterjedését szintén befolyásolhatja, hogy az adott országban milyen közösségi média oldalakat használnak. A Twitter továbbra is nagyon nyitott a kutatók felé adathozzáférési szempontból, tehát azokban az országokban, ahol a Twitter nagyon elterjedt (pl. USA) kisebb a nyomás a kutatókon, hogy más platformokat is monitorozzanak. Az is fontos szempont lehet, hogy az adott ország állampolgárai mennyire adattudatosak és mennyire félnek az adataikkal való visszaéléstől. Ha valahol alacsony az adatmegosztási bizalom, akkor könnyen kudarcba fulladhat egy adatdonációs projekt. A magyar pilotban nem gyűjtöttünk adatot a visszautasításról. A német GESISS kutatóközpontnak viszont vannak ilyen jellegű adatai. A GESIS kutatói egy olyan modellt teszteltek, amiben egy böngésző kiegészítőt kellett telepítenie a résztvevőnek a gépére, ami gyűjtötte, hogy milyen tartalmak jelennek meg a résztvevő Facebook-falán. Ez bár nem DDP, hanem böngésző alapú adatgyűjtés, de ugyanúgy szükséges hozzá a kutatási alany aktív hozzájárulása. Az online Facebookozó minta 60%-a engedélyezte a böngésző kiegészítő feltelepítést, de a gyakorlatban csak 40%-a telepítette az eszközt. Az elutasítók felének voltak adatvédelmi aggályai (Breuer et al. 2021). Felmerülhet alternatív megoldásként, hogy ezek az adatfelvételek ráépülnek olyan online panelekre, ahol már amúgy is magas az adatmegosztási hajlandóság. Ez azonban nem feltétlen segít a donáció alapú adatgyűjtések külső érvényességén.

Az itt felsorolt nehézségek ellenére azt gondolom, hogy a közeljövőben egyre több donáció alapú adatgyűjtés fog indulni. A közösségi média kikerülhetetlen szerepet tölt be a modern társadalmak életében. A társadalomtudomány pedig nem kerülheti meg, hogy vizsgálja azt, hogy ezeken a platformokon mi történik. Ezt pedig akkor tudja legjobban megtenni, ha rálát a felhasználók tényleges tevékenységére. Ebben pedig nagy segítség tud lenni az adatdonációs megközelítés.

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

Irodalom

Boeschoten, L. – Ausloos, J. – Moeller, J. – Araujo, T. – Oberski, D. L. (2020): Digital trace data collection through data donation. arXiv preprint arXiv:2011.09851.

Breuer, J. – Kmetty, Z. – Haim, M. – Stier, S. (2021): User-focused approaches for collecting Facebook data in the "post-API age. Kézirat

Brown, R. C. – Fischer, T. – Goldwich, A. D. – Keller, F. – Young, R. – Plener, P. L. (2018): # cutting: Non-suicidal self-injury (NSSI) on Instagram. Psychological medicine, 48(2), 337-346. (doi.org/10.1017/S0033291717001751)

Bruns, A. (2019): After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. Information, Communication & Society, 22(11), 1544–1566. (doi.org/10.1080/1369118x.2019.1637447)

Csepeli Gy. (2015): Szociológia és a Big Data. Replika, (92-93), 171-176.

Dessewffy T. – Láng L. (2015): Big Data és a társadalomtudományok véletlen találkozása a műtőasztalon. Replika, (92-93), 157-170.

DeVito, M. A. (2017): From editors to algorithms: A values-based approach to understanding story selection in the Facebook news feed. Digital Journalism, 5(6), 753-773. (doi.org/10.1080/21670811.2016.1178592)

Freelon, D. (2018): Computational research in the post-API age. Political Communication, 35, no. 4: 665-668. (doi.org/10.1080/10584609.2018.1477506)

Haim, M. – Nienierza, A. (2019): Computational observation. Computational Communication Research, 1(1), 79-102. (doi.org/10.5117/CCR2019.1.004.HAIM)

Halavais, A. (2019): Overcoming terms of service: a proposal for ethical distributed research. Information, Communication & Society, 22: no. 11, 1567-1581. (doi.org/10.1080/1369118X.2019.1627386)

King, G. – Persily, N. (2020): A new model for industry–academic partnerships. PS: Political Science & Politics, 53(4), 703-709. (doi.org/10.1017/S1049096519001021)

Kmetty Z. (2018): A szociológia helye a Big Data paradigmában és a Big data helye a szociológiában. Magyar Tudomány, 179(5), 683-692.

Kmetty Z. – Bozsonyi, K. (2021): Identifying depression-related behavior on Facebook – an experimental study. Kézirat

Kmetty Z. – Karsai M. – Koltai J. – Kertész J. (2020): Context of burstiness in Facebook activities. IC2S2 Conference, 17-20 July, 2020, poszter. https://fbpilot.tatk.elte.hu/media/e4/18/4f92a3175714df1320bfdfa520c44d9ecd12f61d455866b06af7b23b616a/Burstiness_ic2s2_poster.pdf (utolsó letöltés: 2021. 11. 20.)

ISSN 2063-6415
DOI 10.18392/metsz/2021/3/2

Kmetty Z. – Németh R. (2020): Which is your favorite music genre? A validity comparison of Facebook data and survey data. arXiv preprint arXiv:2002.00501.

Koltai, J. – Kmetty Z. – Bozsonyi K. (2021): From Durkheim to Machine Learning: Finding the Relevant Sociological Content in Depression and Suicide-Related Social Media Discourses. In Pathways Between Social Science and Computational Social Science (pp. 237-258). Springer, Cham. (doi.org/10.1007/978-3-030-54936-7_11)

Lazer, D. – Radford, J. (2017): Data ex machina: introduction to big data. Annual Review of Sociology, 43, 19-39. (doi.org/10.1146/annurev-soc-060116-053457)

Mancosu, M. – Vegetti, F. (2020): What you can scrape and what is right to scrape: A proposal for a tool to collect public Facebook data. Social Media+ Society, 6(3), 2056305120940703. (doi.org/10.1177/2056305120940703)

Moreno, M. A. – Ton, A. – Selkie, E. – Evans, Y. (2016): Secret society 123: Understanding the language of self-harm on Instagram. Journal of Adolescent Health, 58(1), 78-84. (doi.org/10.1016/j.jadohealth.2015.09.015)

Puschmann, C. (2019): An end to the wild west of social media research: A response to Axel Bruns. Information, Communication & Society, 22(11), 1582-1589. (doi.org/10.1080/1369118X.2019.1646300)

Reece, A. G. – Danforth, C. M. (2017): Instagram photos reveal predictive markers of depression. EPJ Data Science, 6(1), 1-12. (doi.org/10.1140/epjds/s13688-017-0110-z)

Ságvári B. (2017): Társadalomtudomány a Big Data korában. Statisztikai szemle, 95(5), 491-504 (doi.org/ 10.20311/stat2017.05.hu0491)

Scharkow, M. (2016): The accuracy of self-reported internet use – A validation study using client log data. Communication Methods and Measures, 10(1), 13-27. (doi.org/10.1080/19312458.2015.1118446)

Shapiro, E. H. – Sugarman, M. – Bermejo, F. – Zuckerman, E. (2021): New approaches to platform data research. Netgain partnership. (last accessed: 15.03.2021) https://www.netgainpartnership.org/resources/2021/2/25/new-approaches-to-platform-data-research (utolsó letöltés: 2021. 11. 20.)

Stier, S. – Breuer, J. – Siegers, P. – Thorson, K. (2020): Integrating survey data and digital trace data: Key issues in developing an emerging field. (doi.org/10.1177/0894439319843669)

Tromble, R. (2021): Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age. Social Media+ Society, 7(1), 2056305121988929. (doi.org/10.1177/2056305121988929)