Régi kéziratok átírása az Amazon „Mechanical Turk” szolgáltatása segítségével

Áttekintés

A gyakran még emberi szemmel is nehezen kibetűzhető kézírásos történelmi dokumentumok átírása lassú és költséges vállalkozás. Ezért van az, hogy például az Egyesült Államok negyedik elnöke, *James Madison* írott hagyatékának csak alig több mint a felét publikálták az elmúlt száz év alatt; a harmadik elnök, *Thomas Jefferson* összegyűjtött munkáinak kiadása, amely 1943-ban kezdődött, pedig még várhatóan további 15 évig fog tartani. A többi irat egyelőre csak mikrofilmről készült digitális másolatok formájában tanulmányozható. Az átírásra szakosodott vállalkozások drágán dolgoznak: 7 és 15 dollár között van az óradíjuk – a szolgáltatás szintjétől függően. Érthető tehát, hogy sok levéltár inkább *crowdsourcing* formájában próbálja meg elvégeztetni az ilyen munkákat. Többnyire saját szervezésben zajlik a dolog: maga az intézmény biztosítja a szoftvert, a webes felületet és a szervert, valamint az utóellenőrzést végző szerkesztőket is. Jól szervezett projekteknél a tömegek bevonása igazán sikeres tud lenni, de az infrastruktúra létrehozása és fenntartása, az informatikai és ellenőri szakemberigény ezt a megoldást is hosszadalmassá és költségessé teheti. Az *Abraham Lincoln* hagyaték szerkesztője egy 2010-es újságcikkben pedig arról panaszkodott, hogy az önkéntes átírók annyi hibát és hiányt produkáltak, hogy ezek kijavítására több idő és pénz megy el, mintha az egészet elölről kezdenék.

Az üzleti és a tudományos világban egyaránt elterjedt megoldás egyes tevékenységek kiszervezése a civil szférába, profi vállalkozók megbízása helyett. A „tömeg” állhat fizetetlen önkéntesekből, de valamilyen módon – például kisösszegű fizetéssel – érdekeltté tett emberekből is. Leginkább az ismétlődő, különösebb szakértelmet nem igénylő, de a számítógépek tudását még meghaladó feladatok (pl. képek osztályozása, adatgyűjtés, kézírás felismerése) esetében lehet jó alternatíva a crowdsourcing. A Zooniverse (*zooniverse.org*) jó példa rá, hogy a civilek is bevonhatók a tudományos munkákba. A portálon 2011-ben már közel félmillió önkéntes regisztrált és dolgozott különböző kutatásokban, a galaxisok alak szerinti osztályozásától (Galaxy Zoo) kezdve, a hajónaplók adatai alapján rekonstruálható korabeli éghajlatig (Old Weather). Az Open Dinosaur Project keretében pedig tudományos publikációkból kellett kigyűjteni a dinoszauruszok végtagméreteit. További hasonló projekteket nagy számban találhatunk a Wikipédiában (*en.wikipedia.org/wiki/List_of_crowdsourcing_projects*).

Kézírások átírására is számos példa létezik már, melyeknél gyakran a közösségi médiát is felhasználják a résztvevők toborzására. A Transcribe Bentham (*ucl.ac.uk/transcribe-bentham*) nevű kezdeményezés célja *Jeremy Bentham*, 18-19. századi angol filozófus kéziratainak átírása. A *US Census Bureau* a régi népszámlálási űrlapokat teszi elérhetővé az interneten, és a reCAPTCHA-hoz (*google.com/recaptcha*) hasonló technológiával vonja be a felhasználókat a kézzel írt adatok felismerésébe (*spie.org/x57241.xml*). Az ilyen átírásos feladatok „tömegesítésére” a Scripto nevű szabad szoftver (*scripto.org*) is jól használható.

A jelen cikkben bemutatott *Written Rummage* projekt szervezői különféle crowdsourcing lehetőségeket és egyéb internetes megoldásokat is számba vettek, amikor *Frederick Douglass*, 19. századi amerikai abolicionista szónok naplójegyzeteit szerették volna kereshető digitális szöveggé alakítani. Az egyik legnehezebb feladat a más projekteknél tapasztalt átírási pontatlanságok kiküszöbölése volt. A megoldás végül az lett, hogy minden dokumentumot kétszer is kiosztottak a bedolgozóknak: a második fázis az első korrektúrájaként szolgált. A munkafolyamatok automatizálása a Google Docs és a Mechanical Turk szolgáltatások segítségével történt.

Megvalósítás

A könyvkereskedőként ismertté vált Amazon cég – *Kempelen Farkas* sakkozógépéről elnevezett – *Mechanical Turk* (röviden: *MTurk*) platformja (*mturk.com*) az egyik legkorábbi internetes crowdsourcing eszköz. A keretrendszer lehetőséget ad arra, hogy különféle feladatokat osszunk ki a tömegnek, meghatározott díjazás fejében (amiből az Amazon is részesedik). Ehhez az első lépés az, hogy *Requester*-ként kell regisztrálkodnunk az MTurk rendszerben, majd a *Design* menüpontot választani. Itt egy úgynevezett HIT (Human Intelligence Task) űrlapot kell kitöltenünk, amit előregyártott sablonok közül választhatunk. Egy átírási feladathoz például meg kell fogalmazni az elvárásokat, be kell linkelni a kézírásos oldalképet és megadni azt a szövegdobozt, ahová a bedolgozó begépelheti a felismert szöveget. Ugyancsak ebben a tervezési fázisban határozhatjuk meg, hogy mennyit kívánunk fizetni egy-egy ilyen elvégzett HIT feladatért. Minél alacsonyabb ez az összeg, annál lassúbb lesz az átírási folyamat, hiszen az MTurk munkásai sokféle egyéb ajánlat közül is választhatnak. A *Written Rummage* projektnél a $0.08/HIT tarifa bizonyult optimálisnak, így egy átlagosan 7 feladatot tartalmazó HIT csomag 3-4 nap alatt készült el.

Az MTurk egyik jó tulajdonsága, hogy egyszerre több feladatspecifikáció is feltölthető egy vesszőkkel elválasztott adatokat tartalmazó (CSV) fájlba, melyben benne vannak az átírandó képekre mutató linkek. A CSV fájl feltöltése és a feladatok beélesítéséhez szükséges néhány végső lépés a *Publish* menüpont alatt tehető meg. A rendszer ezután levonja a számlánkról a szükséges munkadíjat, nekünk pedig már csak várnunk kell. A munka előrehaladását a *Manage* aloldalon követhetjük nyomon és ugyancsak itt találjuk meg az elkészült átírásokat. A Written Rummage projekt adminisztrátorai ezeket egy-egy Google Docs dokumentumba másolták át, majd nyilvánosan megosztották őket. Erre a második, javítási fázis miatt volt szükség, mivel az utóellenőrzést ahelyett, hogy saját maguk végezték volna el (ami szinte ugyanannyi munka lett volna, mintha ők csinálták volna meg az átírást is), szintén az MTurk bedolgozói tömegére bízták. Ám ha csak egyszerűen az eredeti oldalképet és az átírt szöveget kellett volna a korrektoroknak összehasonlítaniuk, akkor nagy lett volna a kísértés arra, hogy egyesek egyszerűen csak átmásolják az első fázisban begépelt szöveget a javítási szövegdobozba és minden ellenőrzés nélkül adják be. Ezért, hogy rákényszerítsék az alapos átolvasásra az ellenőröket, az első átírásban szándékosan hibákat rejtettek el: véletlenszerű betűket írtak egyes szavakba, illetve értelmes szavakat szúrtak néhány mondatba, és ezeket a hibákat egy naplófájlban eltárolták. Ha a korrektor kijavította a szándékos rontásokat, akkor feltehetően az eredeti átíráskor keletkezett egyéb hibákat is észrevette és javította, így megkapta a neki járó díjazást. Ha nem, akkor természetesen nem fogadták el a munkáját és a feladatot újra feltöltötték a Mechanical Turk rendszerébe. A korrektúráért valamivel magasabb tarifát szabtak meg, mint az elsődleges átírásért, tekintve, hogy itt két dolgot: a kézírás oldalképét és a begépelt szöveget kellett gondosan összehasonlítani. Természetesen a két munkafázis párhuzamosan futott: miközben egyes oldalak átírása folyt, a már elkészültek ellenőrzése is zajlott.

Tanulságok

A 72 oldalas napló átírt, kereshető, wiki formátumba szerkesztett szövege – az eredeti kézirat szkennelt oldalképeivel együtt – immár mindenki számára tanulmányozható az interneten (*frederickdouglassdiary.wikispaces.com*). A projektben az időtartam volt a legbizonytalanabb tényező. Kezdetben igen alacsony, 0,01 illetve 0,03 dolláros tarifát állítottak be a két munkafázisra. Ekkor 2-4 hétig kellett várni, mire 5-7 oldal elkészült, úgyhogy különböző változtatásokat tettek mind a munkafolyamatokban, mind pedig a díjazás mértékében. Végül a 0,08, illetve 0,10 dolláros ár bizonyult megfelelőnek (és ehhez jött még az Amazon 10%-os járadéka), így már 6-8 oldal készült el nagyjából hetente. Sőt, ha több ilyen 6-8 oldalas munkacsomagot töltöttek fel egyszerre, akkor sem nőtt a várakozási idő. Az átírással foglalkozó professzionális szolgáltatók oldalanként 2-8 dolláros tarifával dolgoznak, vagyis a napló begépeltetése egy ilyen cégnél valahol 144 és 576 dollár között lett volna. A Mechanical Turk bedolgozói viszont – a kezdeti tesztek költségeit is beleszámítva – mindössze 22,86 dollárért elvégezték a feladatot. Ekkora árkülönbség persze bizonyos etikai kérdéseket is felvet: szabad-e ennyire alacsony munkabérért dolgoztatni embereket? De mivel a crowdsourcing-típusú munkákat sokan önkéntesen, hasznos időtöltésként, díjazás nélkül vagy csak zsebpénz-kiegészítésként végzik, és az MTurk „piacterén” kínált sokféle típusú és tarifájú feladatból mindenki a neki leginkább megfelelőt választhatja, ezért amíg akad vállalkozó egy munka elvégzésére, addig ez egyfajta közös megegye-

zésnek tekinthető az adott HIT elfogadható díjazására vonatkozóan.

A könyvtárak, levéltárak és magánarchívumok egyaránt érzik az igényt, hogy online is elérhetővé tegyék a gyűjteményüket, köztük a kézírásos dokumentumokat is. A digitalizálás egyes unalmas, fáradságos, másként talán el sem végezhető munkafázisaira olcsó megoldást jelenthet a crowdsourcing, vagyis a tömegbe való kiszervezés. A fentiekben ismertetett eljárás nem igényel saját infrastruktúrát és informatikai szakértelmet, ugyanakkor jó minőségű átírt szöveget eredményez néhány hét alatt a legalacsonyabb piaci ár 15 százalékáért. Ezzel a módszerrel a tömeg (munka)erejét felhasználva gyorsabban és nagyobb mennyiségben menthetők át az értékes dokumentumok a digitális világba.

/LANG, Andrew S.I.D – RIO-ROSS, Joshua: Using Amazon Mechanical Turk to Transcribe Historical Handwritten Documents. = The Code4Lib Journal, 15. sz., 2011-10-31
http://journal.code4lib.org/articles/6004/

(Drótos László)

A 10-es lesz az utolsó Windows

A *Microsoft* egyik munkatársa szellőztette meg az információt, amely később megerősítést nyert.

Jerry Nixon az *Ignite Conference* nevű rendezvényen jelentette ki, hogy nagy erővel dolgoznak azon, hogy megjelentessék az operációs rendszert és mivel ez lesz a Windows utolsó verziója, ezért mindannyian részt vesznek a fejlesztésében. Mindez alátámasztja azt, hogy a Windows továbbfejlesztése radikális irányt vehet és az új stratégiában a Windows már nem igazán operációs rendszer, sokkal inkább egy szolgáltatás lesz, amelynél a nagy verzióugrásokat rendszeres frissítések válthatják ki.

Az ötlet egyáltalán nem új, a Microsoft ugyanakkor sosem tért ki arra, hogy miként lehetne az operációs rendszerből szolgáltatás. A stratégia egyik része a rendszerkomponensek felosztása lehet, így például a Start menü akár az operációs rendszer többi részétől függetlenül is frissíthetővé válhat. Ezáltal a társaság az eddiginél sokkal gyorsabban és hatékonyabban reagálhatna a kihívásokra, egyúttal biztosíthatná, hogy a szoftver a különböző platformokon egyaránt optimálisan futhasson.

Nixon szavait megerősítette a Microsoft egyik szóvivője is. Közölte, hogy az elhangzottak azt mutatják be, hogy ők miként akarják szolgáltatásként kínálni a Windowst. Így a felhasználókat és az üzleti ügyfeleiket folyamatosan elláthatják új innovációkkal és frissítésekkel, s a jövőben a verziószámok jelentősége sokkal kisebb lesz. A radikális változásokat már az is mutatta, hogy *Mark Russinovich*, a Microsoft egyik menedzsere áprilisban a *Chefconf 2015* rendezvényen azt mondta, hogy a Windows egyszer nyílt forráskódú lehet, azonban ez nem a közeljövőben fog megvalósulni.

/Forrás: ***https://sg.hu/cikkek/112329/a-10-es-lesz-az-utolso-windows***/

(B. Bné)