2014.11.25.

Észrevételek Mag Kornélia (KSH) az MTA SJTB ülésén elhangzott előadásához

A Big data adatállományokban lévő adatok osztályozása

A nemzetközi Big data dokumentumokban nincs megfelelő definíció az adat/data fogalmára, emiatt a Big data definíciók sem világosak.

Politikai jelentősége a fizikai adathordozón lévő adatoknak, az adatállományoknak van, ugyanis ezeket lehet másolni, feldolgozni, továbbítani, megjeleníteni, összegyűjteni, személyes adattá tenni vagy anonimizálni, menet közben vagy más gépéről, szerverről ellopni, más gépére feltenni, adatbázisba építeni, stb. ezekkel kapcsolatban merül fel a forgalmazás, díjak, adók, a védelem, a jogosulatlan felhasználás, a hozzáférés stb. szabályozása, más döntéshozatal vagy ellenőrzés érdekében a helyzet, a folyamatok elemzése, vagyis kormányzati funkciók.

A KSH által átvett osztályozás ugyanakkor nem a fizikai adatállományok adatút szerinti osztályozását szolgálja, hanem az adatfajták eredeti rendeltetése szerinti osztályozását. A legnagyobb és politikai szempontból legérzékenyebb Big data adatállományoknak éppen az a jellegzetessége, hogy számos különböző forrásból származnak, függetlenül attól, hogy eredetileg ki, milyen célra gyűjtötte azokat. Ezért ez az osztályozás a legfontosabb politikai szempontokból diszfunkcionális. Az e szerint az osztályozás szerint people-to-people adatok sok példányban különféle szolgáltatókhoz, állami szervekhez kerülnek, és azokat ott többször használják fel üzleti és állami célokra, mint a címzetteknél.

Az osztályozás osztályozási ismérvének ilyetén megválasztása miatt a „Közösségi háló (ember által létrehozott információ) – People-to-people tipusú adat" megnevezésű osztályba sorolják a Facebook, Twitter, Tumblr rendszerekbe feltöltött illetve onnan letöltött adatokon kívül a levélíró gépén fogalmazás alatt álló, az éppen úton levő, a szolgáltató gépén lévő, valamint a címzett gépén lévő adatok mindegyikét. Eltekintve attól, hogy a cím három része három különböző dolgot takar, valójában People-to-people adat az interneten, de más távközlési hálózatokon, sőt az egész informatikában sincsen, ezeken csak P2M, M2P és M2M adatok vannak. A people-to-people adat – az információstatisztikában - az egyik ember által a másiknak átadott adat.

Ezzel összhangban én a http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data helyen „human-sourced information"-t találtam, ami korrekt és nem azonos a people-to-people alá tartozókkal.

Másrészt valóban (égető) szükség van az ember által előállított adatok (adat-lábnyom) tényleges bit egységekben mért forgalmának, felhalmozásának, felhalmozott mennyiségének és felhasználásának a megfigyelésére, ugyanis ezek nélkül az adatok nélkül nincs tárgyalási pozíció a nagy globális szolgáltatókkal, nemzetközi szervezetekkel, más államokkal és nem lehet a felhasználókat ráébreszteni arra, hogy kicsiny infokommunikációs tevékenységeikkel milyen folyamatok részei, és tevékenységüknek milyen következményei vannak magukra, az országokra és más országokra nézve, nem lehetnek tudatos felhasználókká.

Érdekes módon a nemzetközi dokumentumokban a telefonbeszélgetések nem szerepelnek sehol, pedig ugye az IP-s beszélgetések ugyanolyan digitális adatok, mint akármi más, mint pl. a beszélő földrajzi helyzete, és ezeket ugyanúgy feldolgozzák, mint a GPS-t, és az e-maileket, amely adatokat a dokumentumok említik.

Big data a KSH-ban

Ezen a KSH szempontjából autentikus helyen http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=99484307 egyébként a Big data-ra ezt a definíciót találtam.

> *Data that is difficult to collect, store or process within the conventional systems of statistical organizations. Either, their volume, velocity, structure or variety requires the adoption of new statistical software processing techniques and/or IT infrastructure to enable cost-effective insights to be made.*

Valóban, a Big data-t nem statisztikai hivatalok és nem statisztikai hivataloknak találták ki, a KSH-ban a személyi állomány jelenleg erre nem is alkalmas. Nem elsősorban állami statisztika készítésére, hanem szolgáltatásnyújtásra és nagy (üzleti, titkosszolgálati, katonai stb.) objektumok és műveletek tervezésére, vezénylésére, menet közbeni ellenőrzésére használatos.

Ha a Hivatalnak annyi baja van, nincs pénze, embere, az adatszolgáltatók nem szolgáltatnak adatot stb. mint mondja, miért nem foglalkozik egyszerűbb dolgokkal, pl. információstatisztikával. Ki van találva, nincs kihívás, jobbára csak el kellene rendelni, feldolgozási költség alig van.

A KSH az ő „kihívásának" viszont a Big data adatállományok megjelenését tekinti. A challenge szép szó, de itthon fogalmazhatunk magyarul is: a KSH jogszabályi felhatalmazást, pénzt, szoftvert, gépeket, hálózatot és megoldásokat: „know-how"-t akar, hogy felhasználhasson valamilyen (milyen is) adatmonstrumokat.

Megoldást, de mire is? Az új féle digitális állam új funkciókat (is) lát el és a régieket is másképp. Vajon a régi mutatórendszer-e a legalkalmasabb egy digitális állam vezetésének támogatására, vagy a rendszer netán fődarabok cseréjére, új egységek beépítésére szorul? Nagy kérdés, kár, hogy e gondolkodás eredménye a KSH statisztikai rendszerén nem szembeötlő.

Miután se pénz, sem munkaerő, se új célok, a Big data adatállományok statisztikai felhasználása a KSH-ban jó eséllyel úgy fog megvalósulni, hogy külföldről adnak pénzt nekünk (mondjuk pilot project, vagy mint régen a Phare program keretében Uszta elvtársunk koordinálásával) a magyar adatok odakint megtervezett begyűjtésére és esetleg feldolgozására, hogy azokat átadhassuk. Így léptünk be az európai hírügynökségbe is: adatot szolgáltatunk, ők integrálják és utána felhasználják a mi integrálásunk érdekében, nekünk pedig nem lesz a felhasználásra pénzünk, emberünk, no meg elképzelésünk.

Nagy általános módszertan és általános módszertani kutatás helyett célszerű lenne a konkrétumokkal foglalkozni. Nem általában foglalkozni a Big data-val, hanem felkutatni az elérhető és esetleg valamilyen égető szükség miatt kiszámítandó mutatók becslésére felhasználható nagy adatállományokat és megvalósíthatósági tanulmányokat készíteni azok felhasználásáról.

Évek óta több alkalommal és helyen is javasoltam, hogy a magyar állam, meghatározott funkciók ellátására, rendszeresítse a távközlési hálózatokban folyó anonimizált adatok statisztikai kiértékelését, amihez persze körültekintés és jó megoldások kellenek, ami egy „kihívás". De a más célra készült adatok országos statisztikai felhasználása már akkor is körültekintést és jó megoldásokat igényelt, amikor Fényes Elek 1838-ban „Ismertető a honi 's külföldi gazdaságban" c. periodikájában rendszeresen közzétette az egyes mezőgazdasági termékek piaci árát. Pl. 14. szám, 168. ill. 29. szám 352. oldal, majd később például a mezőgazdasági statisztika hosszú története során, vagy amikor kialakult a népességnyilvántartás. Persze e feladat megoldása reálisan nem telepíthető a KSH-ba.

A Big data tradicionális statisztikai mérőszámok becslésére történő felhasználásánál pillanatnyilag sokkal nagyobb jelentősége lenne annak, hogy a KSH és a statisztikai hivatalok végre a digitális-kori kormányzás szükségleteinek megfelelően megfigyeljék az adatforgalmat, mint jövőbeni intézkedések és műveletek tárgyát, mert erre viszont minden adottságuk megvan. **Honnan, mennyi magyar adat**

kerül a Big data-kba, egyáltalán az országba, hol vannak és mekkorák ezek az állományok, mennyi magyar adat van bennük, mire használhatók ezek már most, és az egyre újabb funkciókkal hány embert mire lehet majd késztetni, hol vannak a magyar adatok, itthon vagy külföldön, milyenek a belföldi és nemzetközi adatmérlegeink, milyen pénzfolyamok kísérik az adatfolyamokat, ki jár és mennyire jól, stb. stb..

Ahelyett, vagy amellett, hogy pusztán régi mutatókat próbálnak az új adatokkal megbecsülni.

A Big data jelenség és a magyar állam

A Big data adatállományokat azok tudják felhasználni, akiknek vannak olyan feladataik, amelyek megoldásához ez az eszköz, alap.

Azok az államok, amelyeknek vannak/lesznek ilyen adat-erőforrásaik és van/lesz határozott vezetésük, képesek lesznek az ország intézményrendszerét saját (nem „módszertani", hanem valós) új féle módon megfogalmazható céljaik érdekében éppen ezen erőforrásokra építve úgy átalakítani, hogy az képes legyen új funkciókat ellátni, elsősorban új fajta műveleteket tervezni és végrehajtani.

A statisztikai hivatalok, a maguk módszereivel, gondolkodásmódjával a hagyományos államok intézményei, melyek hosszú fejlődésük során mindenkor hagyományos módon működő ipari társadalmi államok adatigényeit elégítették ki.

Nem lenne célszerű a statisztikai hivatalokra, Magyarországon a KSH-ra alapozni az állam új működésmódjait, ezen belül az állam által hozzáférhető adatmonstrumok állandó fenntartásának feladatainak megoldását, mert hiába dolgoztak a régi viszonyok között soknak minősülő adattal, éppen a hagyományos statisztikai hivatalok szigorú módszerei és régi fogalmaikhoz, világképükhöz, munkamódszerükhöz való ragaszkodása miatt.

A szolgálatok gondolkodásmódja, munkamódszere alkalmasabbá teszi őket e feladatokra, itt azonban új funkciók definiálására, jelentős szervezetalakításra és a hatáskörök újrafogalmazásra lesz szükség.

A magyar állam új, az adatmonstrumok köré szerveződő intézményrendszerének koncepcionálását Magyarországon elvben az Államreform Bizottság tudná megoldani, ha nem római jogászokból, a nyugati egyetemi államtudósok tanítványaiból, vagy régi államfajtákban szocializálódott hatalomtechnikusokból hanem olyanokból állna, akik megfelelő tárgyi tudásbirtokában képesek a Christensen szerinti valódi, első tipusú innovációra is.

Részvétel a Big data jelenséggel kapcsolatos nemzetközi tevékenységekben

A magyar statisztikai szolgálatnak kevés a pénze, az embere és kevés és kicsi az általa elérhető adatmonstrum. Ezért már középtávon sem Magyarország lesz itt a tényleges vezető, legfeljebb a jó tanuló stróman.

Az UNECE inventory of big data projects olyan szerepet látszik játszani, mint korábban az afrikai országokba küldött reconnaissance geológusok (v.ö. Rejtő Jenő művei), folyamatosan, in statu nascendi identifikálják a nagy és gazdag országok üzleti és állami szervei által – valamilyen e célra tervezett tranzakciók keretében - a többi országból megszerzendő adatnyersanyagokat.

Ebbe a projektbe a KSH-nak nem kellene annyira belelépni, hacsak mi nem akarjuk megszerezni a Google a Facebook, vagy Koszovó adatbázisait, mert olyan sok pénzünk, emberünk, eszünk és megoldandó feladatunk van a világ népességével vagy Koszovóval kapcsolatban, hogy mást nem is tehetnénk. Nem kell mindenütt elsőnek lenni, erre tanít a Don-kanyar és Auschwitz.

Belföldön a KSH enélkül is meg tudja találni, ami kellhet neki.

Várható, hogy az eddigieken túlmenően is számos nemzetközi ernyőszervezet alakul, illetve nyilvánítja ki érdekeltségét, melynek alapítói a szervezet révén igyekeznek a maguk számára pénzt, adatot, befolyást, tekintélyt, más államok befolyásolási lehetőségét stb. szerezni. Ezt a ME vagy a Külügy koordinálhatná a fentiek tudatában.