ГОЛЕМИТЕ ДАННИ - ВЪЗМОЖНОСТ, ПРЕДИЗВИКАТЕЛСТВО ИЛИ ЗАПЛАХА ПРЕД ОФИЦИАЛНАТА СТАТИСТИКА

Галя Статева[*]

Въведение

В нашия модерен свят все повече данни се генерират от световната интернет мрежа и се произвеждат от електронни сензори и устройства, които са навсякъде около нас. Тези данни се създават в резултат от протичане на процеси в различни сфери на обществено-икономическия живот на национално и международно равнище. Тяхното естество трябва да се разглежда в два основни аспекта: като описание на характеристиките, спецификите и особеностите на всеки процес и като ръководство за управление на всеки процес. Обемът на данните и високата скорост, с която се произвеждат, води до създаване на концепцията **„големи данни“ (Big Data).** Това понятие е добре дефинирано от определението на Gartner[1]:

Големите данни са данните, които могат да бъдат описани като „голямо количество от разнообразни данни, осигуряващи ефективността и ефикасността на протичащите процеси чрез повишаване на познанието и вземането на правилни решения“.

Големите данни се характеризират още като набор от данни с нарастващ обем, скорост и разнообразие, или т.нар. 3Vs[2]. Те обикновено са неструктурирани, нямат предварително дефиниран модел и мащаб, като най-често са в текстови формат.

Статистическите организации дефинират големите данни като:

Данни, които трудно се събират, съхраняват или обработват с конвенционалните системи на статистическите организации. Техният обем,

* Държавен експерт в отдел „Обща методология и анализ на статистическите изследвания”, НСИ; e-mail: gstateva@nsi.bg.

[1] Повече информация може да се намери на адрес: http://www.gartner.com/it-glossary/big-data/.

[2] Volume, Velocity, Variety (3Vs).

скорост, структура и разнообразие изискват адаптиране на нов статистически софтуер за обработка и/или нова ИТ инфраструктура, за да бъдат ефективни направените разходи.

Според потребителите с източниците на големи данни е трудно да се работи. Основните причини за това са: скоростта на промените; трудностите с тяхната идентификация за възможно най-кратък срок от време; големите мащаби; непознаването на методите за тяхната обработка и превръщането им в обобщени синтетични показатели, позволяващи използването им за аналитични цели. Примери за големи данни са: търговски трансакции с кредитни и дебитни карти; данни за околната среда от различни сензори за наблюдение на въздуха, водата и почвата; трафик информация от камери за наблюдение; социална информация от различни мрежи като Twitter, Facebook, Google+, Linkedin и други.

Информационните потоци в обществото се променят динамично. Това обстоятелство поставя нови акценти в общественото развитие. Големите данни създават нови търговски възможности в частния и обществения сектор, но освен това могат да бъдат потенциално интересни и като източник за официалната статистика - за самостоятелно използване или в комбинация с традиционните източници на данни - например извадковите изследвания и/или административните източници. Основателно възникват редица въпроси: как големите данни могат да спомогнат да се измерят точно и навременно икономическите, политическите, социалните и природните феномени в нашия постоянно развиващ се свят?

Големите данни възникват от множество източници, които могат да се групират в три основни категории:

• Население (социални мрежи)

• Данни, генерирани от информационни системи (традиционни бизнес системи и уебсайтове)

• Данни, генерирани от машини/сензорни устройства (автоматизирани системи).

Процесът на „добиване" на информация от големи данни и инкорпорирането им в производствен процес на официалната статистика не е никак лека задача.

Какво се случва, когато официалната статистика срещне големите данни?

Официалната статистика играе ключова роля в модерното общество. До 80-те години на миналия век данните бяха оскъдна стока с висока цена. Преди ерата на големите данни информацията не беше толкова достъпна и трябваше да бъде събирана за определени цели. Статистическите данни се получаваха предимно чрез изчерпателни изследвания. Тези данни се използваха предимно за целите на държавното управление. Такива бяха изследванията за държавните предприятия, селското стопанство и други. През 90-те години изчерпателните изследвания отстъпиха място на извадковите изследвания. Постепенно получаването на статистически данни чрез въпросници беше допълнено и с данни от административни източници. Много статистически служби имат достъп по закон до всички държавни, институционални източници на данни и имат право да събират данни от други източници, без да плащат за това на доставчиците. В

частност, възможността за комбиниране на данни от различни източници прави официалната статистика по-прецизна в своята дейност и разширява нейния обхват.

Във връзка с това и в унисон с високото технологично развитие информацията, осигурявана от статистическите институти, запазва своята уникалност и няма алтернатива, тъй като нейното качество се определя от 15-те принципа на Кодекса на европейската статистическа практика[3]. Професионалистите, работещи в производството на официална статистика, са държавни служители и имат широки и специфични познания в статистическата наука. Те са гарант, който осигурява доверието на обществото в надеждността на произвежданата статистическа информация. Успоредно с това усилията на Европейската статистическа система (ЕСС) са насочени към повишаване на равнището на стандартизиране, хармонизиране и съчетаване на различни източници на данни, включително и на динамично променящите се големи потоци от данни. Това обстоятелство променя облика на официалните статистики в държавите от Европейския съюз.

В контекста на тези обективно случващи се процеси големите данни все повече разширяват своето присъствие в общественото пространство: огромен обем от дигитална информация, произлизаща от всички видове човешки дейности, служи за производство на статистика, която се използва от частни институции или компании. Възникват следните въпроси: дали компаниите, които събират данни, са собственици на тези данни; могат ли да ги използват за различни цели без съгласието на респондентите; дали тези числа могат да се възприемат като конкурентни на официалната статистика; могат ли да се разработват анализи, съчетаващи информация от големи данни и от официалната статистика.

Източниците на големи данни предлагат огромен обем от данни, които изискват съхранение и обработка, надвишаващи капацитета на традиционните статистически средства при процеса на производство на статистика. Поради тази причина биха могли да се прилагат „нови“ техники за извличане на знания от данни (data mining) и прилагане на алгоритми от областта на машинното обучение (machine learning algorithms), имащи изискваната изчислителна ефикасност (Bondi, 2000).

Друго съображение относно използването на големите данни е свързано с представителността и обхвата на произвежданата от тях статистика. С традиционните извадкови техники се осигурява точност на получените статистически оценки на базата на размера на стохастичните грешки. Това може би е приложимо и за големите данни, които могат да бъдат адаптирани към традиционните извадкови техники, но се изискват достатъчно добри аргументи и прецизен анализ на получените емпирични резултати. Успоредно могат да се разработят и алтернативни методи, които отразяват спецификата на големите данни, тяхната динамичност, обхват и области на приложение.

Теоретично погледнато, големите данни могат да бъдат използвани за производство на официална статистика по различни начини: 1) като заменят изцяло статистическите източници, основани на общи дефиниции, класификации и т.н., което е малко вероятно в обозримото бъдеще; 2) частична замяна на

[3] Повече информация може да се намери на адрес: http://www.nsi.bg/sites/default/files/files/pages/Quality/1.1.%20CoP_ALL_BG.pdf.

статистическите източници, като допълват информацията чрез съчетаване на данни от различни източници на данните; 3) осигуряване на напълно нови статистически числа, които могат да допълват и да се интегрират с наличната статистическа информация, което е значително по-добрият начин за тяхното съвместно използване. Първите два начина вероятно биха могли да доведат до намаляване на разходите и натоварването на респондентите, но това, от своя страна, ще доведе до нови задачи за адаптиране, съчетаване и хармонизиране на различни структури от данни към вече утвърдени и общоприети статистически концепции, дефиниции и класификации. Логично погледнато, големите данни не могат да заменят напълно или частично статистическите източници в краткосрочен план и това би било твърде скъпо по отношение на времевите, финансовите и човешките ресурси. Наред с това на този етап от глобализирането на света между статистическите и големите данни се наблюдават моментни процеси на конвергенция, които са необходими за управлението на бизнеса. Фирмите от частния сектор, произвеждащи статистика на основата на големите данни, следват третия път и не се сблъскват с подобни проблеми.

Предизвикателства пред официалната статистика

Срещата на големите данни с официалната статистика води до много и различни предизвикателства.

Едно от най-големите предизвикателства пред статистиците при използването на големите данни засяга **методологията**. При провеждане на едно традиционно изследване статистиците дефинират генерална съвкупност, разработват дизайн на извадката, събират данните и т.н. Много източници на големи данни като например съобщенията в социалните мрежи нямат добре дефинирана генерална съвкупност, структура и качество. Това прави трудно прилагането на традиционните статистически методи, основани на теорията на извадковите изследвания. Основната особеност е, че при работа с големи данни първо идват събраните данни и след това статистиците трябва да приложат специфични методи за тяхната обработка и анализ (визуализационни методи, техники за извличане на знания от данни или други методи, които да „направят големите данни малки"). Наред с методологията важен елемент на официалната статистика е осигуряването на информация с високо **качество.** Дали сегашните принципи и стандарти за качество са пряко приложими за големите данни, или е необходимо да бъдат подходящо адаптирани?

Поверителността и правните въпроси са друго предизвикателство. Защитата от разкриване на идентичността на индивидите е задължителна, но това е трудно да се осигури, когато е свързано с големите данни. Проблемът с големите данни е, че често потребителите на услуги и устройства, генериращи данни, не са запознати, че правят това и/или за какво тези данни могат да бъдат използвани впоследствие. Друг правен въпрос е свързан с авторските права и собствеността на данните. Дори ако данните могат легално да бъдат използвани, това не предполага че е разумно или подходящо да се направи с оглед нарушаване на тяхната конфиденциалност. Например в някои случаи може да бъде полезно да се приложи подходът на информираното съгласие. Някои договори за абонамент на мобилна

услуга включват клауза за използване на данни от договора за други цели освен за предоставяне на самата услуга.

Обработката, съхранението и трансферът на големи масиви от данни създават предпоставки за възникване на чисто **технологично** предизвикателство. Технологичният напредък - нарастването на компютърната мощност, по-големи ИТ устройства за съхранение и високоскоростни канали за данни, може частично да реши тези въпроси. Събирането на данни в реално време отваря нови възможности за комбиниране на административни данни с високоскоростни големи данни, идващи от различни източници, като търговски данни (трансакции с кредитни карти, онлайн разплащания, продажби и др.), мобилни устройства и сензори (мобилни телефони, GPS, камери, метеорологични сензори, сензори за замърсяване на въздуха и др.), социални медии (Twitter, Facebook, Google) и други обществено достъпни данни.

Друго предизвикателство е **възможната променливост и съпоставимост** на източниците на големи данни, имайки предвид факта, че официалната статистика поддържа анализ на динамичните редове. За много потребители продължаването на тези динамични редове е от особена важност и това не може да бъде пренебрегнато.

Не на последно място е размерът на **финансовите ресурси,** които статистическите институти трябва да заделят, за да придобият права на собственост върху големи данни, чиито собственици са компании от частния сектор. Още повече, че съгласно настоящото законодателство статистическите институти придобиват данни от държавни институции, вкл. от административни регистри и респонденти, безвъзмездно.

За да използва ефективно големите данни, официалната статистика се нуждае от експерти с **различно мислене и нови умения,** които да могат да извличат ценно „познание" от данните. Такива специалисти са т.нар. изследователи на данни („data scientists").

Големите данни за официалната статистика означават и по-голям обем информация, която е предмет на дефиниране на **нови политики и директиви** за управление и защита на тази информация.

Всички изброени предизвикателства пораждат основателния въпрос: не е ли дошло време за големите данни и тяхното място в официалната статистика да се мисли неконвенционално и в нова перспектива? Очевидно е, че към момента мисленето изостава по отношение на информационните технологии и големите данни.

Методите на статистиката и големите данни

Възниква фундаменталният въпрос: могат ли големите данни да се обработват и анализират с методите на статистиката? По-конкретно, това означава изследователите да могат да използват корелационния, дисперсионния и хи квадрат анализ, когато проверяват своите научни тези. В допълнение това означава още, че те могат да се възползват от целия инструментариум за проверка на хипотези. Очевидно е, че към традиционния инструментариум от теорията на статистиката ще бъдат включени и други подходи за изследване и анализ. Това обстоятелство на практика откроява две насоки за развитие на научноизследователската работа:

• Използване на добре известните статистически методи за анализ. Това създава условия за съчетаване на данните от официално провежданите статистически изследвания с потоците от големи данни.

• Въвеждане на други подходи, начини и методи за самостоятелно или съчетано използване на големите данни.

Възниква също въпросът: какво печели официалната статистика от използването на големите данни за аналитични цели? Очакванията в тази посока са най-вече в две направления:

• Намаляване на натоварването на респондентите. Това е особено важно, тъй като в последните години статистическите изследвания се развиват особено екстензивно. Потребностите от информация нарастват лавинообразно. Нарастват също и изискванията по отношение на качеството, бързината и детайлизацията на информационния продукт. Все повече внимание се обръща на информационната осигуреност на малки териториални пространства. Стремежът да се дефинират най-добрите и оптимални управленски решения налага разработването на нови показатели и подходи за съчетаване на данните от текущата статистика и големите данни.

• Намаляване на цената на информационния продукт. Известно е, че статистическите изследвания имат своя цена, която се повишава в зависимост от поставените условия за представителност, точност и достоверност на статистическите оценки. Колкото по-високи са изискванията в това отношение, толкова по-големи ще бъдат разходите за направените изследвания. Стойностите могат да бъдат екстремални. Например ако се постави изискването точността на оценките от едно извадково статистическо изследване да нарасне два пъти, то обемът на извадката трябва да нарасне четири пъти. Това означава, че разходите за изследването (отпечатване на документи, заплати, транспортни разходи, консумативи, обучение на по-голям брой анкетьори, евентуално заплащане на респондентите и т.н.) също нарастват четири пъти при равни други условия.

Най-големите опасности от технологична гледна точка, които следва да се имат предвид, когато се съчетават данните на официалната статистика и големите данни, са:

• Достоверност на големите данни. Разгледани в този аспект, големите данни изключително много зависят от източника на тяхното генериране. Използването на индикатори, гарантиращи достоверността на източниците, е от първостепенно значение за качеството на големите данни.

• Представителност на големите данни. От съществено значение за представителността на големите данни е тяхното селектиране, избор и систематизация. Това означава, че субективният подход за тяхното генериране трябва да бъде елиминиран. Принципите и условията на рандомизация следва да бъдат водещи при определяне на данните, които се използват в съчетание със статистическите данни, отговарящи на същите условия.

• Точност на големите данни. Обемът на големите данни трябва да бъде достатъчен, за да се редуцира стохастичната грешка. Това означава, че големите данни имат висока аналитична значимост. Тази предпоставка предопределя качеството на изводите и посланията към потребителите.

На практика това са фундаментални понятия от теорията на извадковите статистически изследвания, които следва да бъдат ключът към успешно, целево и аналитично съчетаване на данните от официалната статистика и големите данни.

Как големите данни могат да бъдат използвани в статистическия бизнес процес?

Използването на източниците на големи данни в реалното производство на статистика е все още предмет на предварителни проучвания. Някои европейски страни вече са стартирали задълбочени проучвания в тази насока, такъв пример е Нидерландия по отношение на **статистиката от социалните медии.** Социалните медии са потенциален източник на големи данни. В тях хората доброволно споделят информация, дискутират интересуващи ги теми и общуват със семействата и приятелите си. Около един милион съобщения от социалните медии се генерират ежедневно в Нидерландия и са достъпни за всеки, който използва интернет. Нидерландският статистически институт изучава тези съобщения от две гледни точки: като съдържание и като мнение. Изучаването на съдържанието на съобщенията напр. в Twitter показва, че 50% от тях са с безсмислено съдържание и не носят никаква информация. Останалите са свързани предимно с дейностите през свободното време (10%), работата (7%), телевизията и радиото (5%) и политиката (3%). Изразяването на субективно мнение в социалните мрежи разкрива интересни потенциални възможности на тези източници за статистически цели. Мненията са тясно корелирани с потребителското доверие и в частност с мненията по отношение на икономическата ситуация. Оказва се, че тази корелационна връзка е стабилна на месечна и седмична база. Ежедневните данни не са толкова надеждни, тъй като показват висок дял на променливо потребителско поведение. Всичко това доказва, че е възможно производството на седмични и/или месечни статистически индикатори за потребителското доверие за първия работен ден след наблюдаваната седмица, като се отчита възможността за постигане на бързи резултати.

Успоредно с националната практика Евростат също стартира някои инициативи по отношение на големите данни. В годишните работни програми на Евростат са дефинирани някои предварителни проучвания и изучаването на възможности за използване на източниците на големи данни в следните статистически области:

Статистика на цените - използване и анализ на цените, събирани от интернет. Това е 24-месечен проект (стартирал през януари 2013 г.) за разработване на специфичен софтуер, подпомагащ специалистите при автоматизирано събиране на цените от интернет за изчисляване на индекс на потребителските цени. След приключване на проекта софтуерът ще бъде тестван методологично и технологично в пет европейски страни. Софтуерът, който ще бъде разработен в рамките на проекта, ще бъде предоставен и на други статистически организации за работа под лиценз (EUPL).

Статистика на туризма - предварително проучване за използване на мобилно позиционирани данни за статистиката на туризма. През януари 2013 г. стартира 15-месечен проект, който изследва резултатите от използването на мобилно позиционирани данни за статистиката на туризма (и свързаните с него области) и оценява предимствата и недостатъците. Въпросите, които са предмет на проекта, са достъпът (и непрекъснатостта на достъпа), доверието (на

производителите и потребителите на статистика), разходите, понятията (транслиране на съществуващите концепции за статистиката на туризма към новите източници на данни) и други методологични въпроси (например представителност и обем на извадката и др.). Възможността за обхващане на големи файлове от данни, генерирани от мобилните оператори, се счита за предизвикателството, което трябва да се преодолее в бъдеще като следствие от успешно завършилия проект.

Използване на ИКТ - предварително проучване за използването на трафика на интернет потоците за събиране на статистика за информационното общество. Чрез този проект Евростат цели да изучи пилотно и да оцени приложимостта на потребителски ориентирани и уебориентирани измерителни подходи от техническа, методологическа, разходна, правна и социално-политическа перспектива. В рамките на проекта са планирани три отчета, които са ценни за практиката на националните статистически институти: 1) Как да се разработи официална процедура за събиране на статистика за използването на ИКТ от източници на големи данни; 2) Методологично и технологично ръководство за внедряване; 3) Тестване на концепцията „обединени отворени данни“[4], отнасяща се за споделен набор от големи данни от частния сектор, които ще бъдат отворени за използване от статистическите институти.

През периода 2015 - 2020 г. се очаква развитието на множество проекти с цел проучване и използване на големите данни като достоверен и надежден източник за производство на официална статистика.

Бъдещи стъпки - сътрудничество: с кого и как?

Изправяйки се пред тези предизвикателства, официалната статистика ясно осъзнава необходимостта да не работи в изолация, а в сътрудничество с всички останали заинтересовани страни. Целта на това сътрудничество се състои обикновено в обмяна на знания, опит и добри практики.

Един възможен и необходим партньор са **потенциалните доставчици на големи данни:** ако те не гарантират достъпа до техните данни заради правилата за поверителност, историята ще приключи, преди да е започнала. Но тъй като източниците на големи данни не са пригодени за статистическо използване, такова сътрудничество е от съществено значение, за да се постигне добро познание за произхода на тези източници. Освен това за статистическото производство ще бъде по-ефективно да обработва данните на мястото на тяхното събиране и съхранение, т.е. данните няма да бъдат осигурени безплатно. От друга страна, статистиците могат да извършват анализи на големите данни, което ще помогне на техните собственици да ги разбират и тълкуват по-добре. По този начин взаимоотношенията с доставчиците на данни биха могли да преминат в истинско и ползотворно сътрудничество от взаимен интерес. В допълнение, официалната статистика може да играе и специфична роля на доверителна трета страна. На пазара конкурентите ще бъдат недоверчиви да споделят чувствителни данни един с друг. Но те могат да споделят тези данни със статистическите институти, събиращи статистическа информация, която е от полза за всички. Водещото правило в тази ситуация е създаване на статистическа култура и статистическо мислене. Работата в тази посока ще доведе до покриване на белите полета между експертите.

[4] „Federated open data”.

Сътрудничеството между официалната статистика и **академичната общност** може да се разшири по отношение на големите данни - например при решаването на методологични проблеми, разработване на технически решения и обучение на бъдещи изследователи на данни (data scientists**)**. В дългосрочен план новите знания и умения за големите данни ще изискват адаптиране на университетските учебни програми и планове за обучение по нови дисциплини - например за изследователи на данни, или надграждане на дисциплините в специалност „Статистика" (някои европейски университети вече предлагат подобни курсове на своите студенти). В краткосрочен и средносрочен план е необходимо статистическите институти да предложат на своя експертен състав специализирани обучения по темата за големите данни и по този начин да изградят вътрешен аналитичен капацитет. Международното сътрудничество в тази насока ще бъде много полезно за общността на официалната статистика. Такова сътрудничество също може да бъде подкрепено от разработването и финансирането на различни проекти, които подпомагат научноизследователската и иновативната дейност. Натрупването на знания е ключов момент за постигане на кохерентност между потоците от информация.

Има много **търговски партньори от частния сектор,** с които статистическите институти биха могли да си сътрудничат. Социалните мрежи Google и Facebook са два ясни примера, за които големите данни формират сърцевината на техния бизнес модел. Тяхното познание и данните, до които имат достъп, могат да бъдат и съотносими за официалната статистика. ИТ компаниите също притежават съответстващи знания за обработката, съхранението, сигурността и облачните технологии на големите данни.

Евростат и Европейската комисия са предприели различни инициативи в рамките на ЕСС, имащи за цел да изследват пълния потенциал на големите данни за европейската икономика, общество и обществени услуги.

Други международни организации (ОИСР, Статистическата комисия на ООН, Световната банка, UNECE) също са въвлечени в развитието на темата за големите данни. Показателен пример за това е проектът на стратегическата международна група по отношение на ролята на големите данни в модернизирането на статистическия производствен процес. Целта на проекта е да направи възможно статистическите организации да действат единно чрез стандартизиран подход по методологически въпроси, свързани с големите данни и производството на статистика, да демонстрират приложимост на използването на източниците на големи данни за официална статистика и да подпомагат обмяната на знания, експертиза и средства между статистическите организации.

Някои **национални статистически офиси** вече осъществяват проекти и са запознати добре с големите данни. Техният опит от успешното използване на големите данни може да бъде изучаван и споделян с други страни с цел извличане на ценни познания и прилагане на добри практики по отношение на източниците на големи данни. Освен това националните статистически организации са окуражавани да включат официално въпросите за големите данни в техните годишни програми и стратегически документи чрез осъществяване на изследователски и пилотни проекти в избрани области и чрез разпределяне на подходящите ресурси за тези цели.

Интегриране на големите данни в ЕСС

За да посрещне предизвикателствата и да създаде подходящи условия за сътрудничество между частните единици и държавните статистически институти за достъп до нови източници на данни, Евростат е разработил пътна карта и план за действие. Всички заинтересовани страни от политиката, науката, статистиката и бизнеса, както и обществото като цяло, ще бъдат привлечени.

Целта на пътната карта е да направи възможно постепенното интегриране на източниците на големи данни в производството на европейска и национална статистика в дългосрочен, средносрочен и краткосрочен аспект, като по този начин ще допринесе значително за постигане на целите на „Визия 2020" на ЕСС.

В дългосрочната визия (след 2020 г.) е залегнала идеята, че източниците на големи данни трябва да бъдат окончателно интегрирани в производствения бизнес процес на официалната статистика в ЕСС. Очаква се след 2020 г. източниците на големи данни да бъдат налични за широк кръг от статистически продукти и съответното законодателство да претърпи съществени промени, така че да позволи използване на източници на големи данни за официална статистика и да се работи в условия, когато обществото ще „признае" използването на тези източници за статистически цели.

Средносрочните цели (до 2020 г.) включват пълна интеграция на стратегията за големите данни в националните стратегии, интегриране на подходящи ИТ инфраструктури, методологии, избрани приложения в статистическото производство, публично-частно партньорство с доставчиците на данни за целите на официалната статистика и увеличаване на броя на статистиците, занимаващи се с големи данни в ЕСС.

Краткосрочните перспективи (до края на 2016 г.) са насочени към дейности, свързани предимно с анализиране и подготовка на условия и инфраструктура за използването на големите данни в официалната статистика и стартиране на конкретни пилотни проекти за придобиване на знания и опит. Тези проекти включват дейности по анализ на източниците на големи данни, изследване на потенциала за партньорство с доставчиците на данни, развитие на обучителни програми, идентифициране на изследователските нужди, интегриране на стратегията за големите данни в цялостната стратегия на европейско ниво, доразвиване и внедряване на комуникационна стратегия.

Планът за действие е свързан с хоризонталните дейности, както и с провеждането на пилотни проекти. Целта на пилотните проекти е да се добият специфични умения при използване на големите данни в контекста на официалната статистика. Това би могло да включва идентифициране, анализиране и решаване на проблеми от хоризонтално естество или теми, както и обследване и развитие на бъдещи бизнес модели за производството на статистически данни, свързани със специфични източници на данни или статистически продукти. Предложеният подход изисква координация между пилотните проекти, свързани с източниците на данни и дейностите по теми. Евростат идентифицира следните групи по теми:

• Политика

Стратегията за големите данни в официалната статистика трябва да бъде вградена в управленските стратегии на национално и европейско ниво. Големите

данни трябва да бъдат част от осигуряването на информация за социално-икономическите програми за управление.

• Комуникация

Някои от източниците на големи данни съдържат чувствителна информация. Използването на тези източници за целите на официалната статистика може да предизвика негативни възприятия в обществото и заинтересованите лица. Ето защо е необходимо да се дефинират целите, процедурите и резултатите от използването на големи данни съгласно Кодекса на европейската статистическа практика, като фокусът е върху етичните принципи - например поверителността.

Наред с това в процеса на комуникация се верифицират данните от различните източници на информация, включително и информацията на официалната статистика.

Трябва да се има предвид, че медиите са ключов посредник между статистическите офиси и широката общественост и включването им в цялостната комуникационна стратегия е от съществено значение за успеха на плана за действие.

• Източници на големите данни

Броят на източниците на големи данни нараства лавинообразно. Разнообразието и размерът им определят и разширяването на потенциала на големите данни за производството на официална статистика. За да се вземат информирани решения за действие и за пилотните проекти, е важно да се работи за създаването и разработването на наръчник и за класифициране на източниците на големите данни. По този начин ще се осъществява мониторинг по отношение на тяхната автентичност, качество и достоверност.

• Приложения/пилотни проекти

На европейско и международно ниво вече са предприети действия по отношение на големите данни - например създаването на Task Force група в Евростат и High Level група на ниво Статистическа комисия на ООН. С тяхната дейност стартират процеси, които ще се разклоняват във времето и пространството.

• Методи

Използването на източници на големи данни изисква приложение на нови методи при анализа и обработката на данни. В същото време методите са зависими от източниците на данни, например дали тези източници съдържат структурирани данни, или просто текстова информация. Очевидна е необходимостта от ревизия на досега използваните методи за обработка и анализ на информацията, както и от нови подходи в тази насока. Това е продължителен процес с основна цел да се постигне ефективност при разкриването на закономерности в информационните потоци, формиращи големите данни, което е от първостепенно значение за доброто управление и натрупването на знания.

• Качество

Осигуряването на информация с високо качество е един от най-важните елементи на официалната статистика. Статистическата информация, която ще се произвежда от източниците на големи данни, е необходимо да е пригодна и да отговаря на стандартите за качество на ЕСС, за да бъде използваема от потребителите. Това е гаранция за доброто интегриране на данните от различни източници.

• ИТ инфраструктура

Характеристиките на големите данни, включително техният обем, скорост и разнообразие, оказват влияние върху ИТ системите и инфраструктурата. Бъдещите ИТ инфраструктури ще бъдат определени предимно от новите бизнес модели, които ще се внедрят за производство на статистика от големите данни. Това е процес, който е характерен за глобалната икономика.

• Умения

Достъпът, управлението, обработката и анализът на големите данни изисква специфични нови умения или комбинация от умения, които към настоящия момент не са част от официалната статистика. Тези нови умения се идентифицират с понятието „изследовател на данни" („data scientist") и стремежът е да има все повече такива специалисти в официалните статистически организации. Наред с това успехът на аналитичното мислене повече от всякога ще зависи от умението да се екстрахират най-важните данни от много източници за кратко време, като се има предвид динамиката на изменение в информационните потоци. Уникалността на експертите в това отношение ще зависи от тяхната интуиция и сензитивност към информационните единици.

• Обмяна на опит

Важен елемент е споделянето на опит за проекти, приложения, пилотни проекти и източници на големи данни в ЕСС. Обмяната на добри практики е важен момент за ускоряване на процесите по опознаване и рационално използване на големите данни.

• Законодателство

Законодателството е от основно значение при определянето на законова рамка за достъп, обработка и разпространение на статистически продукти от източници на големи данни. Това означава, че то трябва да бъде насочено към осигуряване на среда, където доминират точно установен ред, регламенти и информационна защита. Сложността на законодателството при тези обстоятелства зависи и от определянето на правилата, по които функционират информационните системи, контролирани и използвани от различни органи на управление.

• Управление

Планът за действие и пътната карта са интегрална и неделима част от „Визия 2020" на ЕСС. Дейностите, свързани с големите данни, са ключова област, дефинирана в Плана за внедряване на „Визия 2020". Предизвикателствата при използването на големите данни на практика са неделима част от „Визия 2020". Традиционните методи за управление, независимо дали вършат добра работа, се нуждаят от обновление. Успешното управление се осъществява единствено и само въз основа на знание. Въпросът е как това знание да се добие от информационните потоци, циркулиращи в общественото пространство.

Заключение

Бъдещето на официалната статистика в ерата на големите данни все още е въпрос на обсъждане и експериментиране. Необходимо е цялата международна и европейска статистическа общност да се адаптира към новата реалност и да отговори адекватно на възможностите и предизвикателствата, които тази реалност предлага. За да се случи това, е нужно широко и ползотворно сътрудничество с различни „играчи" във и извън статистическата общност чрез формиране на

разнообразни мрежи, които могат да изградят нови пътища за генериране на статистически данни. Това е гаранция за успешно обществено развитие в условията на глобалната икономика.

За всички, които се занимават със статистика, ерата на големите данни е много вълнуващо време. Време за натрупване на нови знания с поглед към бъдещето. В обозримия хоризонт информационните потоци ще играят все по-осезаема роля в управлението на световните пазари. Технологията за използване на големите данни е в началния етап на своето развитие. Все още се дефинират рамките, мащабът и обхватът на дейностите. Това, което със сигурност се знае, е, че процесите са обективни и необратими. Контролът и управлението са невъзможни без качественото използване на потоците информация, които се генерират лавинообразно.

ЦИТИРАНА ЛИТЕРАТУРА:

Big data - an opportunity or a threat to official statistics? Paper prepared by Eurostat for the April 2014 plenary session of the Conference of European Statisticians as part of the seminar entitled „What is the value of official statistics and how do we communicate that value?“ (http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=100307471).

Big Data (and official statistics). Piet Daas and Mark van der Loo, Statistics Netherlands.

ESS Big Data Action Plan and Roadmap 1.0 Work Programme. Eurostat paper - room document.

Official statistics and Big Data, Peter Struijs, Barteld Braaksma and Piet JH Daas Big Data & Society 2014 1: DOI: 10.1177/2053951714538417.

What does „Big Data" mean for official statistics? Michael Glasson (Australia), Julie Trepanier (Canada), Vincenzo Patruno (Italy), Piet Daas (Netherlands), Michail Skaliotis (Eurostat) and Anjum Khan (UNECE) (http://www1.unece.org/stat/platform/display/hlgbas).