УДК 316.34:311.3-048.35(477+100)

ВЕЛИКІ ДАНІ (BIG DATA) І МОДЕРНІЗАЦІЯ НАЦІОНАЛЬНИХ СИСТЕМ ОФІЦІЙНОЇ СТАТИСТИКИ

О. Л. Єршова,
кандидат економічних наук, доцент,
в. о. завідувача кафедри інформаційних систем і технологій;
Т. В. Томашевська,
кандидат технічних наук,
доцент кафедри інформаційних систем і технологій;
Національна академія статистики, обліку та аудиту

Анотація. У статті розглянуто питання, пов'язані з генерацією великих обсягів неструктурованих даних різної природи (великі дані, або Big Data). Проаналізовано можливості використання Big Data в офіційній статистиці, обговорюються їх джерела і проблеми при переході статистичних служб на їх використання, а також необхідність розширення сфер використання Big Data в Україні.

Аннотация. В статье рассмотрены вопросы, связанные с генерацией больших объемов неструктурированных данных разной природы (большие данные, или Big Data). Проанализированы возможности использования Big Data в официальной статистике, обсуждаются их источники и проблемы при переходе статистических служб на их использование, а также необходимость расширения областей использования Big Data в Украине.

Abstract. The article deals with issues related to the generation of large amounts of unstructured data of different nature (big data). The basic features of the Big Data are investigated. The authors analyzed the possibility of using Big Data in the official statistics, their sources and discussed problems in the transition of statistical services to their use. We also discuss the need of the areas of Big Data usage in Ukraine expansion and attention attraction to the issue of statistical services.

Big Data may potentially provide more relevant and timely statistical data training for official statistics then the traditional sources. Including the sources of Big Data into its preparation process of official statistics, national, regional and international statistical organizations could strengthen their position in terms of a more timely and cost-efficient production of official statistics by economic sectors, social and environmental costs with fewer resources. Big Data may help to improve the challenges of timely preparation of harmonized statistics on the economic, social and environmental decision-making purposes, research and public debate.

Dealing with Big Data suggests the modernization of the statistical system. It will be necessary to strengthen the statistical system research sector to increase information technology capabilities and restructure the human resources.

Постановка проблеми. В результаті всесвітнього використання електронних пристроїв і повсюдного виробництва цифрової інформації кардинально змінився характер даних, які генеруються зараз постійно й у величезних кількостях – це так звані великі дані (Big Data). Ці дані мають важливі властивості, які відрізняють їх від даних, отримуваних із традиційних джерел. Вони мають широкий діапазон розподілу, неорганізовану структуру, великий обсяг і часто надходять в масштабі реального часу.

В епоху, коли знижується частка респондентів, які беруть участь в обстеженні домашніх господарств і підприємств, великі дані можуть забезпечувати директивні органи фактологічною інформацією в масштабі реального часу в таких сферах, як ціни, зайнятість, обсяг виробництва, економічний розвиток і динаміка населення. Великі дані можуть потенційно забезпечувати підготовку більш актуальних і своєчасних статистичних даних порівняно з традиційними джерелами для офіційної статистики. Включивши джерела великих даних у процес підготовки офіційної статистики, національні, регіональні та міжнародні статистичні організації могли б зміцнити свої позиції у частині своєчасного й економічно ефективнішого отримання даних офіційної статистики за секторами економіки, соціальної сфери і навколишнього середовища з меншими витратами ресурсів. У зв'язку з цим Генеральний секретар Організації Об'єднаних Націй оголосив у 2009 році про реалізацію ініціативи "Глобальний пульс". Її мета – упровадження науково-технічних нововведень у галузі цифрової інформації і швидкого збирання й аналізу даних, щоб керівники, які приймають рішення, могли в масштабі реального часу отримувати повне уявлення про те, як кризи впливають на стан вразливих верств населення.

Аналіз останніх досліджень і публікацій. Незважаючи на те, що термін Big Data міцно увійшов у корпоративну мову, вони залишаються майже недослідженим українськими вченими. Серед зарубіжних учених слід зазначити публікації К. Лінча [1], В. Майєр-Шенбергера та К. Кук'єра [2], Ж.-П. Дейкса [3], які дослідили сутність, типи та принципи Big Data. Корисними є також дослідження вчених Варшавського інституту економіки С. Бухгольца, М. Буковські, А. Шнегольські, які проаналізували вплив Big Data на європейську економіку [4]. У контексті вивчення підходів до застосування в дослідженнях концепції Big Data важливими є наукові праці К. Лінча, який вперше увів це поняття в публікації "Як можуть вплинути на майбутнє науки технології, що відкривають можливості для роботи з великою кількістю даних?" у спецвипуску журналу "Nature" 3 вересня 2008 року [1, с. 28]. Незважаючи на певну розробленість проблеми Big Data зарубіжними вченими ряд аспектів їх застосування залишається малодослідженим. Потенційні можливості їх застосування для статистичного аналізу глобальних явищ у країнах, регіонах та в усьому світі є перспективним напрямом досліджень.

Метою дослідження є аналіз аспектів застосування Big Data для офіційної статистики в масштабах усього світу та перспективи їх використання в Україні.

Виклад основного матеріалу. Уряди дедалі більшої кількості країн визнають важливість великих даних та створюють спільноти фахівців-практиків і робочі групи для вивчення питання щодо їх використання та отримання від них потенційної віддачі. Статистичне співтовариство поступово усвідомлює, що у цій сфері назріває якісний стрибок.

На своїй другій нараді, що відбулася 21–22 жовтня 2013 року, Бюро Конференції європейських статистиків 2013/2014 років, що є керівним органом Європейської економічної комісії (ЄЕК) в галузі статистики, провело поглиблене вивчення проблематики великих даних [5]. За його результатами сформульовано такі основні рекомендації: міжнародному статистичному співтовариству слід спільними зусиллями визначити ключові пріоритетні галузі використання великих даних і взятися за їх освоєння; слід створити механізм для обміну інформацією щодо знань та досвіду використання великих даних. Бюро схвалило також проект, присвячений великим даним, який має такі цілі:

а) виявити основні можливості, які передбачають великі дані, підготувати методичні вказівки для статистичних організацій, а також розробити скоординовані заходи з вирішення основних стратегічних і методологічних питань, які виникають в секторі офіційної статистики у зв'язку з використанням Big Data;

б) продемонструвати доцільність ефективної підготовки як нової статистичної продукції, так і традиційної офіційної статистики з використанням джерел великих даних і можливість тиражування цих підходів у різних національних умовах;

в) сприяти обміну знаннями, технічним досвідом, інструментами і методами між організаціями з метою підготовки статистики з використанням джерел великих даних.

Джерела виникнення Big Data для офіційної статистики можна класифікувати таким чином:

• джерела даних, пов'язані зі здійсненням програми, будь то державної або іншої (наприклад, електронні медичні картки, відомості прийому клієнтів лікарняними установами, облікові страхові документи, облікові банківські документи тощо);

• комерційні або операційні джерела даних, пов'язані зі здійсненням операцій між двома сторонами (наприклад, операції з кредитними картками, онлайнові операції, в тому числі здійснювані за допомогою мобільних пристроїв);

• джерела даних, пов'язані з роботою сенсорних мереж (наприклад, дані із зображень, отриманих зі супутників, дані з автодорожніх датчиків і метеорологічні дані з вимірювальних пристроїв);

• джерела даних, пов'язані з роботою пристроїв реєстрації (наприклад, реєстрація даних з мережі мобільного телефонного зв'язку і з Глобальної системи визначення координат (GPS));

• джерела даних, пов'язані з поведінкою користувачів (наприклад, дані пошуку в Інтернеті того чи іншого продукту, послуги або будь-якого іншого виду інформації, дані про перегляди веб-сторінок);

• джерела даних, пов'язані з висловленням користувачами своїх думок (наприклад, дані щодо коментарів у соціальних мережах).

Дані з адміністративних документів є одним з головних джерел інформації для підготовки офіційної статистики національними статистичними системами. Такі дані, отримувані від державно-адміністративних органів, традиційно мають сильно структурований характер і потім обробляються, зберігаються, систематизуються й використовуються статистичними відомствами. Адміністративні документи сьогодні не є джерелом великих даних, однак вони можуть стати ним у разі збільшення швидкості приросту і фізичного обсягу, наприклад коли статистичні відомства почнуть ширше користуватися даними з адміністративних документів завдяки їх збиранню в масштабі реального часу або на щоденній чи щотижневій основі, а не раз на рік або раз на місяць, як це зазвичай робиться нині.

Застосування великих даних в офіційній статистиці породжує багато проблем, серед яких:

• юридичні, тобто пов'язані з доступом до даних і їх використанням;

• пов'язані з недоторканністю приватного життя, тобто з використанням громадської довіри й отриманням згоди на вторинне використання даних і їх ув'язку з іншими джерелами;

• фінансові, пов'язані з потенційними витратами на вилучення даних із джерела порівняно з отримуваними вигодами;

• управлінські, пов'язані, наприклад, із політикою та директивами з питань управління даними і забезпечення їх захисту;

• методологічні, пов'язані з якістю даних і придатністю статистичних методів;

• технологічні, пов'язані з інформаційними технологіями.

Взявши за основу визначення, наведені в доповіді цільової групи ЄЕК, Статистичний відділ розробив анкету, присвячену використанню великих даних для підготовки офіційної статистики. Анкета складається з трьох основних частин: джерела, проблеми і сфери використання Big Data. Це всесвітнє оцінювання було проведено з метою отримання інформації про національні пріоритети, події і досвід в частині поточного або планованого використання великих даних для підготовки офіційної статистики.

Анкету було розіслано статистичним відомствам більш ніж 200 країн і територій у період з липня по вересень 2013 року. Анкета розсилалася англійською мовою в липні, іспанською – в серпні та французькою – у вересні. Її можна було заповнювати або через Інтернет, або в документі у форматі pdf. Станом на 2 листопада 2013 року було отримано 107 відповідей. У повному вигляді отримані результати представлені Статистичною комісією як довідковий документ. В узагальненому вигляді ці результати викладені нижче.

Питання, що стосується джерел великих даних, було сформульовано так: "Будь ласка, вкажіть, які із зазначених нижче джерел великих даних будуть, ймовірно, використовуватися протягом наступних 12 місяців Вашим

управлінням або іншими установами, які є частиною національної статистичної системи Вашої країни".

У разі позитивної відповіді респондентам пропонувалося пояснити, які конкретні джерела даних вони збираються використовувати. Більше 50% країн і територій повідомили, що вони збираються використовувати адміністративні джерела в якості джерел великих даних, а для кожного з інших п'яти джерел даних показник використання склав близько 25%. Кілька країн підняли питання про те, є об'ємні адміністративні документи джерелом великих даних чи ні. Дані з адміністративних джерел є основою для підготовки багатьох видів статистичної продукції, однак питання про те, чи слід розглядати їх у тому самому контексті, що і великі дані, потребує обговорення.

Питання, що належить до розділу "Проблеми, пов'язані з використанням великих даних", було сформульовано так: "Чи становить [те чи інше питання] серйозну проблему для національної статистичної системи у Вашій країні?". Було запропоновано такі можливі відповіді: "Ні" (не викликає особливих проблем), "Не маю думки з цього питання" (це питання не обговорювалося) або "Так" (є проблемою). За всіма шістьма категоріями проблем більшість країн відповіли "Так" (є проблемою), наступною за частотністю була відповідь "Не маю думки з цього питання" (тобто ці питання ще не були предметом ретельного обговорення), і лише невелика кількість країн відповіли "Ні". В цілому найчастіше вказувалися методологічні, інформаційно-технологічні та управлінські проблеми, за якими з невеликим відставанням слідували юридичні проблеми та проблеми, пов'язані з недоторканністю приватного життя.

У третій, заключній частині всесвітнього обстеження респондентам пропонувалося вказати сфери використання (або вивчення питання про використання) великих даних протягом наступних 12 місяців. На вибір було запропоновано такі галузі офіційної статистики: "Демографічна та соціальна статистика", "Статистика природного руху населення і записів актів громадянського стану", "Економічна і фінансова статистика", "Статистика цін", "Статистика транспорту", "Статистика навколишнього середовища" та ін. Респондентам було дано два варіанти відповіді: "Ні" або "Так", при цьому відповідь "Так" потрібно було супроводжувати поясненням.

Серед галузей, в яких використовуються великі дані, найчастіше вказувалися "Демографічна та соціальна статистика" , а також "Економічна і фінансова статистика". Однак, як і в разі використання джерел великих даних, позитивні відповіді в цій частині обстеження необхідно було повторно проаналізувати, виділивши ту групу відповідей, яку можна включити до категорії "належної практики". Після проведення такого повторного аналізу виявилося, що належна практика використання великих даних має місце в галузі статистики цін, економічної і фінансової статистики. Зокрема, кілька країн вказали, що вони використовують дані сканування та / або методи просіювання веб-сторінок для розрахунку часто оновлюваних індексів цін, які використовуються на додаток до

стандартного індексу споживчих цін. У цілому належна практика використання великих даних в галузях статистики становить трохи більше 10% в галузі статистики цін, економічної і фінансової статистики, демографічної і соціальної статистики та близько 5% у кожній з інших галузей.

Результати дослідження [4] показують, що до 2020 року Big Data можуть збільшити ВВП країн Європи на 1,9%, що є еквівалентом річного зростання в ЄС. Побудована макроекономічна модель дозволила вченим спрогнозувати очікуваний економічний ефект за секторами економіки, а саме: зростання до рівня 23% і 22% в торгівлі і виробництві відповідно; по 13% – у сферах державного управління і фінансових послуг; на 6% – у секторі інформаційно-комунікаційних технологій; на 5% – в охороні здоров'я та на 19% – в інших секторах.

Більшою мірою ефект Big Data залежатиме від ефективності управління і використання ресурсів. Країни з великими підприємствами, глобальними зв'язками, з розвинутою інфраструктурою інформаційно-комунікаційних технологій зможуть отримати значно більшу вигоду, ніж ті, що в цьому відстають. Таким чином, наслідки від використання великих даних будуть більш відчутними в Північній Європі, водночас у більшості країн нової Європи і Південної Європи результат буде значно меншим.

В Україні застосування великих даних лише на слуху у вузьких професійних колах ІТ-індустрії. Проводяться конференції, тренінги та професійні курси здебільшого у напрямі комерційного застосування Big Data. Розроблення продуктів із застосуванням Big Data здійснюється для іноземних замовників, в той час як ця можливість стала окремим бізнесом в світовій ІТ-індустрії. Конференції, які відбуваються в Україні на цю тему, показують, що основними напрямами застосування Big Data є задачі в галузях маркетингу, електронної комерції, у банківській справі та телекомунікаціях. Про можливості та необхідність їх застосування в офіційній статистиці мова не йде. Тобто українські фахівці (статистики та ІТ-розробники) цим питанням не цікавляться.

Висновки з дослідження і перспективи подальших розвідок у цьому напрямі:

1. Потенційні можливості використання великих даних для підготовки офіційної статистики визнані статистичною спільнотою. Big Data можуть допомогти більш ефективному виконанню завдання своєчасної підготовки узгоджених статистичних даних про економіку, соціальну сферу та екологію для прийняття рішень, проведення досліджень і громадських обговорень. Крім того, з удосконаленням технології геокодування статистичних одиниць в економічній, соціальній та екологічній сферах і збільшенням обсягу наявної інформації на найнижчому рівні географічної деталізації очікується якнайшвидший прогрес у галузі використання великих даних у розвинених країнах і країнах, що розвиваються. Тому для того, щоб скористатися перевагами Big Data, між регіональними ініціативами всередині світового статистичного співтовариства необхідно налагодити обмін методичними

розробками, кращими практиками вирішення стратегічних питань і можливостями навчання, в тому числі в справі вирішення питань, пов'язаних із законодавчою базою, недоторканністю приватного життя, фінансами, управлінням, методологією та технологіями.

2. Вирішення проблем, пов'язаних із використанням великих даних, передбачає модернізацію статистичної системи. Необхідно зміцнити дослідний сектор статистичної системи, наростити інформаційно-технологічні можливості, провести структурну реорганізацію людських ресурсів шляхом залучення вчених з галузі інформатики та налагодити партнерські відносини з приватним сектором у сфері обміну автоматично генерованою інформацією в цифровому форматі при дотриманні принципів недоторканності приватної життя, правил конфіденційності.

3. Великі дані та модернізація статистичних систем створюють для більшості країн дуже схожі проблеми і можливості. Тому між національними статистичними відомствами може і має бути налагоджений аналогічний обмін досвідом, практичними методами і рішеннями.

4. Українським фахівцям відповідних галузей (статистика та IT) варто зацікавитися можливістю некомерційного застосування Big Data. Для впровадження останніх в офіційну статистику повинна бути розроблена державна програма та передбачено розв'язання комплексу проблем: юридичних, технічних та технологічних. Перехід на застосування Big Data в офіційній статистиці сприятиме інтеграції української статистики у світову статистичну спільноту.

Список використаних джерел

1. Lynch C. How do your data grow? / C. Lynch // Nature. – 2008. – V. 455, No 7209. – P. 28–29.

2. Шенбергер В. М. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / В. М. Шенбергер, К. Кукьер ; пер. с англ. И. Гайдюк. –М. : Манн, Иванов и Фербер, 2014. – 240 с.

3. Dijcks J.-P. Big Data for the Enterprise [Electronic resource] /J.-P. Dijcks // Oracle. – October, 2011. – Access mode : http://bigdatawithoracle– 521307.pdf

4. Buchholtz S. Big & Open Data in Europe: A grow then ginevra missed opportunity? / S. Buchholtz, M. Bukowski, A. Sniegoski // Report commissioned by demosEUROPA – Centre for European Strategy Foundation within the "Innovation and entrepreneurship" programme. – Warsaw, Mdruk, 2014. – 116 p.

5. Conference of European statisticians. Second Meeting of the 2013/2014 Bureau Geneva (Switzerland), 21–22 October 2013. In-depth review of Big Data [Electronic resource]. – Access mode : ECE/CES/BUR/2013/OCT/2