ИНФОРМАЦИЯ ЗА ESSnet ПРОЕКТ „BIG DATA[1]”

*Галя Статева**

През ноември 2015 г. Националният статистически институт се включи като страна партньор в ESSnet проект „Рамково споразумение за сътрудничество **Big Data План за действие”,** който е разделен на две отделни грантови споразумения (SGA-I и SGA-II) и ще се изпълнява последователно за периода 2016 - 2018 година. В рамките на това споразумение ще се постигат целите, заложени в т.нар. BDAR (Big Data Action Plan and Roadmap v. 1.0), който е част от портфолиото на ECC Vision 2020.

Продължителността на SGA-I проекта е 18 месеца (февруари 2016 - юли 2017 г.), на обща стойност 1.111.111 евро и в него ще вземат участие консорциум от 20 национални статистически института и 2 органа на статистиката, като координатор на проекта е Статистическият институт на Нидерландия. Заради мащабността и значимостта на дейностите проектът е организиран в следните девет работни пакета (WPs) по видове източници на Big Data за генериране на статистически данни на европейско ниво:

WP 0: Координация

WP 1: Извличане на информация от интернет за свободни работни места (web scraping)

WP 2: Извличане на информация от интернет за характеристики на предприятията (web

[1] Терминът „Big Data” се използва едновременно на английски и български език и възможно най-адекватният български еквивалент е „големи данни” (към момента не е намерен по-точен превод на понятието). Английският термин придобива все по-голяма популярност в различни български източници.

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

scraping)

WP 3: Данни от измервателни устройства за потребление на електричество

WP 4: Данни от автоматични идентификационни системи (AIS-data)

WP 5: Данни от мобилни телефони

WP 6: Комбинирани източници на данни за краткосрочна бизнес статистика

WP 7: Комбинирани източници на данни за структурна бизнес статистика

WP 8: Методология за събиране, обработка и анализ на Big Data

WP 9: Разпространение и популяризиране на резултатите от проекта

Националният статистически институт е активен партньор в работните пакети **WP2, WP8 и WP9.**

Основната цел на проекта е да подготви Европейската статистическа система (ЕСС) за интегриране на източници на „големи данни“ в процеса на производство на официална статистика. Специфичните цели на работните пакети са свързани с анализиране на получените резултати от източниците на Big Data; разработване на методология за използването на Big Data в статистическата практика и измерване на качеството на събраните данни; идентифициране, дефиниране и внедряване на ИТ инфраструктура за обработка и съхранение на Big Data; правни въпроси, свързани с достъпа и използването на източниците на Big Data в рамките на ЕСС; обмен на информация между официалната статистическата система и научната общност. По същество работата на всеки от петте пилотни пакета (WP 1 - WP 5) се разделя на няколко етапа: осигуряване на достъп до „големите данни“, вкл. писмени партньорства с държателите на тези данни; обработка на Big Data при спазване на тяхната методология; получаване на статистически изходи и оценка на тяхната приложимост.

Накратко съдържанието и очакваните резултати от работните пакети в грантово споразумение SGA-I са, както следва:

- **Извличане на информация от интернет за свободни работни места (web scraping)**

Основната цел на пакета е извличане и тестване на данни от интернет сайтове за търсене и предлагане на работни места Основното предизвикателство е анализ на качеството на получената информация (напр. вид на информацията по отношение на наименованието на длъжността, професията, икономически сектор, работно място и др.) и проучване на възможността за използването им в регулярното изследване на статистиката на труда. Като един от най-важните очаквани резултати по този работен

пакет се очаква да бъдат разработени техники за събиране на данните и правните аспекти за достъп до вече „извлечени“ подобни данни от частни фирми.

- **Данни от измервателни устройства за потребление на електричество**

Основната цел на пакета е разработване на методология и ИТ инфраструктура за тези големи масиви от данни, за да се направи оценка на потреблението на домакинствата и предприятията по NACE (ниво раздел). Като допълнение ще бъде направена оценка на съотношението на свободните жилищни пространства или сезонните свободни жилищни пространства. Критичната точка на този пакет е наличието само на един набор от данни - за Естония.

- **Данни от автоматични идентификационни системи (AIS-data)**

Основната цел на пакета е събиране на AIS данни от кораби и разработване на методология за тях, включително визуализиране на резултатите. Като допълнителни резултати ще бъде генерирана оценка на броя на посещенията на пристанища и разработване на отчетна форма за тях, както и изучаване как тези оценки могат да бъдат свързани с административните регистри на пристанищната администрация. По този начин ще се даде възможност за последващо изучаване (в SGA-II) на възможността дали отчетната форма може да повиши качеството на оценките за транспортираните товари, сравнимостта на тези оценки между страните, както и възможността за разработване на нови статистически продукти - например статистика за трафика на кораби от дадени координационни точки за определени времеви интервали.

- **Данни от мобилни телефони**

Основната цел на пакета е разработване на бизнес план за достъпа до данни от доставчици на мобилни данни и поверителността на клиентите на мобилните оператори. Възможните хипотези за бъдещи пилотни проекти са: оценка на вариациите в съвкупността за обществените потребности и последваща необходимост за връзка между пилотните проекти и „изискуемите“ почистени микроданни от националните мобилни оператори. Същинската част на събиране на данни от мобилните оператори ще се осъществява в SGA-II.

- **Комбинирани източници на данни за краткосрочната бизнес статистика**

Основната цел на пакета е проучване на комбинирани източници на „Big Data", административни и други налични източници на данни за целите на краткосрочната бизнес статистика. За източниците, които имат най-добър потенциал, ще бъде изготвен бизнес план и предложение за включване в SGA-II. В настоящото грантово споразумение ще бъде тествана методологията, ИТ инфраструктурата и качеството на данните за два пилотни проекта: 1) Експресни прогнози за индексите на оборота - бързо производство на показателите за оборот (t+15, t+30,..), чрез уеббазирани анкети за продажби във възможно най-кратки срокове. Като допълнителни източници биха могли да се използват финансови променливи и стойността на заплати и допълнителни възнаграждения; 2) Индексът на потребителското доверие (CCI) е показател, който се дефинира като степен на оптимизъм за състоянието на икономиката, което потребителите изразяват чрез техните дейности да спестяват и харчат. За събиране на данни от източници на „Big Data" ще бъдат използвани социалните мрежи (Twiter, Facebook, http://newsfeed.ijs.si/,...), за да се замени традиционният подход на изчисляване на CCI чрез създаване на експресни оценки и по-надеждни резултати.

- **Комбинирани източници на данни за структурната бизнес статистика**

Основната цел на пакета е да се намери адекватен начин как да се комбинират източници на „Big Data", административни данни и статистически данни от регулярните изследвания, които да обогатят крайния статистически продукт в следните области: „Население", „Туризъм" и „Селско стопанство". В много случаи един източник на данни не е достатъчен за производство на надеждна официална статистика и се налага комбинирането на няколко различни източници на данни. Този работен пакет има научноизследователски характер и е необходимо от методологична гледна точка да се работи с висока степен на професионална независимост.

- **Разпространение и популяризиране на резултатите от проекта**

Основната цел на пакета е да се изгради подходяща уебплатформа за споделяне на състоянието на текущите дейности по проекта и публикуване на окончателните резултати.

- **Извличане на информация от интернет за характеристики на предприятията (web scraping)**

Този работен пакет е първият, по който НСИ вече започна същинска работа. Основната цел е проучване на възможността дали чрез комбинация от различни техники (web scraping, text mining и machine learning) за извличане на данни от уебсайтове може да се получи достатъчно значима и съдържателна информация за характеристиките на предприятията. Тази цел може да бъде постигната на различни нива: на микрониво, за да се обогати и допълни информацията за предприятията в бизнес регистрите; на макро-ниво, за да се произвеждат статистически продукти. Предизвикателството е да бъде оценено качеството (точността) на резултатите както на микро-, така и на макрониво. Пакетът е разделен на четири задачи, като всяка от тях описва конкретни дейности за изпълнение:

Задача 1 - Достъп до Big Data

- Изготвяне на списък от предприятия, които ще бъдат обект на web scraping: а) дефиниране на съвкупност от предприятия според техните характеристики (икономическа дейност и размер); б) идентифициране на бизнес регистри, съдържащи информация, необходима за избор на предприятията.
- Идентифициране на URL адреси за предприятията от списъка: а) подбор на архиви, съдържащи URL адреси, свързани с предприятията, включени в съвкупността на изучаване и оценка на обхвата; б) в случай на недостатъчен обхват, разработване на приложения за търсене на уебадреси на предприятията по идентификатори (напр. наименование, фискален код, икономическа дейност и т.н.) и оценка на надеждността на получените URL.
- Проучване (на европейско и национално ниво) на правните аспекти и въпросите за поверителността на данните вследствие на „извличане“ на адреси на уеб-сайтове.

Задача 2 - Боравене с данните

- Подробно дефиниране на случаи на използване: провеждане на консултации с потребители и други заинтересовани лица за анализиране на нуждите от статистически данни. Тази задача включва координация с дейностите по проект ESS.VIP „Европейска система за оперативно съвместими бизнес регистри”.

- Преглед на научната литературата относно техники и наличните свободни софтуерни продукти за масивен web scraping (JSoup, HTTrack, и т.н.) и изучаване на проблемите, свързани с достъпността на уебсайтовете (блокиращ механизъм). Внедряване на една или повече web scraping техники с цел „извличане“ на съдържанието на уебсайтовете на предприятията.
- Провеждане на същинското „извличане“ на съдържание на уебсайтовете на предприятията във всяка страна участничка в работния пакет.
- Съхранение на събраното съдържание от уебсайтовете в база данни, за да бъдат споделени с всички партньори по проекта (евентуално в ИТ средството UNECE Sandbox, за да бъдат оценени и от правна гледна точка).

Задача 3 - Изпитване на методи и техники

- Избор на някои случаи на използване (извън предварително дефинираните в задача 2, които да осигуряват добра представителност на общите потенциални статистически продукти и информация, за да се обогатят бизнес регистрите.
- Изграждане на доказателства на избраните случаи на използване, за да се прогнозират характеристики на предприятията чрез прилагане на техники за извличане на текст и данни.

Задача 4 - Финализиране на методи и техники

- Избор на извадка от уебсайтове и ръчно определяне на свързани характеристики на предприятия и/или използване на резултатите от изследването „Използване на ИКТ от предприятията”, за да се валидират някои характеристики на предприятията.
- Прилагане на техники за извличане на текст и данни, за да се прогнозират характеристики на предприятията. Оценяване на „прогнозираните“ характеристики на предприятията чрез показатели за качество (напр. точност, чувствителност, специфичност). Сравняване и евентуално интегриране на получената информация с информацията от националните бизнес регистри.

В рамките на проекта **през февруари 2017 г. в София** ще бъде организиран заключителен семинар, на който **НСИ ще бъде домакин и ще участва активно в неговата организация и провеждане.**