

ДОКЛАД ГЛОБАЛЬНОЙ РАБОЧЕЙ ГРУППЫ ПО ВОПРОСАМ ИСПОЛЬЗОВАНИЯ БОЛЬШИХ ДАННЫХ ДЛЯ ПОДГОТОВКИ ОФИЦИАЛЬНОЙ СТАТИСТИКИ*

Большие данные – одна из магистральных тем в развитии мировой статистики сегодня. Журнал «Вопросы статистики» открывает серию публикаций одним из ключевых материалов, разработанных по решению Статистической комиссии ООН. «В докладе излагаются основные итоги работы Международной конференции по вопросам использования больших данных для подготовки официальной статистики, решения первого совещания Глобальной рабочей группы и результаты обследования по вопросу об использовании больших данных для подготовки официальной статистики. В докладе сформулированы круг ведения и предлагаемая программа работы Глобальной рабочей группы, охватывающая учебную и профессиональную подготовку и наращивание потенциала; увязку больших данных с целями в области устойчивого развития; агитационно-просветительскую и коммуникационную деятельность; доступ к данным и механизмы партнерства; и такие сквозные вопросы, как классификации и платформы; а также изучение конкретных источников больших данных для подготовки официальной статистики, а именно данных из сети мобильной телефонной связи, данных из социальных Интернет-сетей и изображений, полученных со спутников. В заключительном разделе доклада изложены вопросы для обсуждения»¹.

Введение. Вследствие широкого распространения и постоянного использования телекоммуникационных и иных устройств, стимулируемого научно-техническим прогрессом, происходит непрерывное генерирование цифровой информации, например данных с устройств Глобальной системы определения координат (GPS), из банковских автоматов, со сканирующих устройств, из сенсорных сетей, из сети мобильной телефонной связи, со спутников или из социальных Интернет-сетей. Такие данные огромного объема, скорости прироста и многообразия, требующие новых инструментов и методов сбора, систематизации и обработки на эффективной основе, обычно называются *большими данными* и могут быть потенциально полезными для подготовки официальной статистики. Однако в целях осуществления капиталовложений в эту новаторскую отрасль статистики статистическим управлениям многих стран еще предстоит подготовить технико-экономические обоснования, в которых излагались бы предпосылки и направления полезного и актуального использования больших данных. Обнадеживающими в этой связи явились итоги Международной конференции по вопросам использования больших данных для подготовки официальной статистики, которые подробнее рассматриваются ниже, поскольку на Конференции

были приведены наглядные примеры многих текущих проектов в области использования больших данных.

Потенциал источников больших данных заключается в возможности своевременного получения из них иногда в режиме реального времени больших массивов данных, которые обычно генерируются с минимальными затратами. Проведение мероприятий, связанных с такими традиционными источниками данных, как обследования домашних хозяйств и предприятий, требует времени на проведение и часто сопряжено со значительными финансовыми затратами на этапе подготовки данных. Собираемые таким образом данные могут быть дополнены, сокращены или заменены благодаря использованию больших данных. Однако статистическое сообщество знает тот факт, что оно сможет воспользоваться преимуществами этих принципиально новых источников данных, в том числе применять их для контроля за достижением целей в области устойчивого развития и представления соответствующей отчетности лишь в том случае, если ему удастся должным образом решить вопросы, касающиеся методологии, качества, технологии, доступа к данным, законодательства, соблюдения неприкосновенности частной жизни, управления и финансирования, и представить надлежащий анализ затрат и результатов, прежде чем большие дан-

* Доклад был представлен на 46-й сессии Статистической комиссии ООН (3-6 марта 2015 г.); публикуется с сокращениями. URL: http://www.un.org/ga/search/view_doc.asp?symbol=E/CN.3/2015/4&Lang=R

¹ Из сопроводительной записки при представлении доклада на 46-й сессии Статистической комиссии ООН.

ные можно будет использовать для подготовки официальной статистики.

В этой связи Статистическая комиссия на своей 45-й сессии, состоявшейся 4-7 марта 2014 г., признала, что большие данные представляют собой источник информации, который нельзя игнорировать и который необходимо оценить по достоинству. Для этого Комиссия поддержала предложение о создании Глобальной рабочей группы по вопросам использования больших данных для подготовки официальной статистики, которая составит перечень уже осуществляемых мероприятий и примеров использования больших данных, рассмотрит вопросы, связанные с методологией, людскими ресурсами, качеством и конфиденциальностью, и разработает руководящие принципы классификации различных видов источников больших данных.

Статистическая комиссия подчеркнула, что в основу круга ведения и мандата Глобальной рабочей группы должны быть положены стратегические соображения, в частности связи с программой развития на период после 2015 г., инициативой в области информационной революции и Основопологающими принципами официальной статистики. Комиссия просила также Группу дополнять и опираться на работу, проводимую региональными комиссиями и другими международными статистическими учреждениями, и уделять особое внимание в своей программе работы конкретным обстоятельствам, в которых находятся развивающиеся страны, например существующим нормативно-правовым базам или некоторым недостаткам имеющейся информационно-технологической инфраструктуры.

Глобальная рабочая группа по вопросам использования больших данных для подготовки официальной статистики была создана в мае 2014 г. во исполнение решения 45/110 Статистической комиссии и провела свое первое совещание 31 октября 2014 г. в Пекине, сразу же по завершении Международной конференции по вопросам использования больших данных для подготовки официальной статистики, совместно организованной в Пекине 28-30 октября 2014 г. Статистическим отделом Департамента по экономическим и социальным вопросам Организации Объединенных Наций и Национальным статистичес-

ким управлением Китая. Кроме того, Статистический отдел и Европейская экономическая комиссия (ЕЭК) провели совместное исследование проектов использования больших данных и организационных условий их осуществления в качестве вклада в подготовку к этим совещаниям.

В следующем разделе дается общий обзор результатов Международной конференции по вопросам использования больших данных для подготовки официальной статистики, а затем представлены краткие итоги работы первого совещания Глобальной рабочей группы и результаты обследования проектов использования больших данных и организационных условий их осуществления. В заключительном разделе доклада излагаются вопросы для обсуждения.

Международная конференция по вопросам использования больших данных для подготовки официальной статистики. На Конференции присутствовало около 120 участников из более чем 40 стран, представляющих различные региональные и международные организации, а также исследовательские учреждения, научные круги и частный сектор. В центре внимания участников конференции были три группы источников больших данных, а именно:

- a) сеть мобильной телефонной связи и устройства GPS и другие регистрирующие устройства;
- b) изображения, полученные со спутников, и другие виды геопространственной информации;
- c) сеть «Twitter» и другие социальные Интернет-сети.

Участники Конференции обсудили типологию каждого из этих источников, конкретные связанные с ними проблемы и механизмы партнерства, необходимые для их использования. На заключительных заседаниях конференции обсуждались общие преимущества и проблемы, связанные с источниками больших данных, пути внедрения новшеств и порядок подготовки технико-экономических обоснований для использования больших данных.

Значительное число представленных на Конференции проектов свидетельствовали о духе новаторства, существующем внутри статистического сообщества. Применение различных источников больших данных в целом

ряде отраслей статистики показывает, что большие данные обладают реальным потенциалом в плане совершенствования официальной статистики. Эти новаторские инструменты могут дополнить собой инструментарий статистической системы, расширяя имеющиеся виды применения статистических данных или обеспечивая поиск более гибких и краткосрочных путей решения весьма актуальных вопросов политики. Различные проекты показали также, как можно решить некоторые из проблем, связанных с методологией, доступом к данным, соблюдением неприкосновенности частной жизни и профессиональной подготовкой. В последующих пунктах изложены некоторые основные моменты работы Конференции.

Во всех технических подробностях был обсужден вопрос о данных из сети мобильной телефонной связи, в том числе различия между активными и пассивными данными о местоположении абонента мобильного телефона. Было рассказано о ряде представляющих интерес видов применения таких данных в статистических целях и опыте некоторых стран, в частности в области статистики туризма и мобильности населения в дневное время и для оценки данных переписей населения, составления карт распространения нищеты и отслеживания картины мобильности населения на случай эпидемических вспышек. В условиях повсеместного распространения мобильных устройств, в том числе в странах развивающегося мира, данные из сети мобильной телефонной связи обладают большим потенциалом в плане предоставления недорогостоящей информации в режиме реального времени по актуальным вопросам развития. Основные проблемы по-прежнему связаны с сохранением конфиденциальности, доступом к данным из сети мобильной телефонной связи и доверием общественности к использованию таких данных.

Изображения, получаемые со спутников, обладают большим потенциалом, особенно при подготовке статистики сельского хозяйства, в плане предоставления более частых и более своевременных данных с высокой степенью дезагрегирования; однако методы оценки, например для оценки урожайности сельскохозяйственных культур, все еще находятся на этапе проверки. Текущая работа Ав-

стралийского статистического бюро в области использования данных с изображений, полученных со спутников, может дополнить и частично заменить нынешнюю практику обследований, проводимых для статистического измерения масштабов сельскохозяйственного производства. Получение изображений со спутников происходит с частотностью один раз в две недели и может привести к снижению частотности (и соответствующей стоимости) проведения обследований. Изучение аналогичных видов использования изображений, получаемых со спутников, в области официальной статистики ведется в Китае, Колумбии и Мексике, а также в других странах, в том числе в целях проверки и экспериментального применения в сфере экосистемного учета.

Поскольку в сетях «Twitter», «Facebook» и других социальных Интернет-сетях, являющихся источниками данных, содержится, по утверждениям многих, крупнейший массив данных о поведении людей, статистическое сообщество уже занялось изучением этих источников на предмет их использования в таких отраслях социологии, требующих больших объемов данных, как, например, вопросы, связанные с охраной здоровья. Проект в Нидерландах явился весьма многообещающим примером того, как оценки настроения потребителей, получаемые на основе данных из сетей «Facebook» и «Twitter», могут в потенциале служить источником предварительных оценок и снизить необходимость проведения обследований, поскольку оценки на основе данных из социальных сетей могут быть получены с большей частотностью и с меньшими затратами. В Италии и Китае используются инструменты «прочесывания» Интернета для получения оценочных показателей норм вакансий, которые могут служить основой при подготовке текущей статистики труда благодаря обеспечению усовершенствованных ежемесячных прогнозов и более точных территориальных оценок. Статистическое сообщество добивается значительно прогресса в этой области, даже несмотря на наличие многочисленных методологических проблем.

Обсуждались также вопросы схожести источников данных не только с точки зрения методологии и качества, но и применитель-

но к соблюдению неприкосновенности частной жизни, механизмам партнерства и информационным технологиям. Было признано, что каждый из различных источников больших данных так или иначе сопряжен со всем комплексом проблем и что ознакомление с этими проблемами во всех деталях является насущной необходимостью. Как только проблемы, с которыми сопряжено использование источников больших данных, будут лучше осмыслены, статистическое сообщество сможет эффективнее использовать их общие черты. Например, вне зависимости от конкретного источника данных могут быть разработаны типовые модели всеобъемлющих соглашений о доступе к данным с компаниями, действующими на мировой арене. Статистическому сообществу следует не конкурировать, а взаимодействовать с частным сектором в целях освоения потенциала больших данных для подготовки официальной статистики. В то же время ему следует оставаться беспристрастным и независимым и вкладывать средства в рекламирование преимуществ задействования богатого потенциала имеющихся цифровых данных на благо общества. Формирование доверия общественности станет ключом к успеху.

Наконец, обсуждались текущие и будущие функции сообщества официальных статистиков. Например, официальным статистикам, возможно, придется чаще реагировать на просьбы о подтверждении информации, поступающей с рынков. Было подчеркнуто, что статистическому сообществу следует по-прежнему четко помнить возложенную на него миссию, а именно предоставлять высококачественную и беспристрастную информацию по актуальным вопросам обществу в целом, и государственным директивным органам в частности. Хотя внедрение технических новшеств является решающим фактором и статистическому сообществу следует не отставать от этого процесса, традиционная работа по основным направлениям будет продолжаться, и необходимость в такой работе будет сохраняться. Например, результаты применения больших данных необходимо будет сопоставлять с контрольными параметрами, а для определения этих параметров необходимо будет проводить традиционные обследования.

Статистическому сообществу следует подготовить и составить технико-экономические

обоснования для использования больших данных. Следует позаботиться о том, чтобы использование больших данных было увязано с программой развития на период после 2015 г. и с целями в области устойчивого развития, которые будут в ней сформулированы, поскольку работа статистических систем в предстоящие годы будет во многом определяться этими целями. Разумеется, не все множество задач, предлагаемых в программе развития на период после 2015 г., может измеряться с помощью показателей, основанных на больших данных. Ожидается, однако, что большие данные сыграют важную роль благодаря своей своевременности и геопространственной детализации. Одним из направлений дальнейшей работы является изучение возможности применения аппроксимативных показателей, основанных на больших данных, которые могли бы обеспечить более частое поступление информации по сравнению с обследованиями. Таким образом, большие данные дополняют, а не подменяют собой традиционные системы. Тем самым они схожи, например, с «мгновенными» оценками валового внутреннего продукта (ВВП).

В целом, участники Международной конференции по вопросам использования больших данных для подготовки официальной статистики провели общий обзор перспективных видов применения больших данных для подготовки официальной статистики и способствовали лучшему пониманию методологических проблем и проблем, связанных с соблюдением неприкосновенности частной жизни и доступом к данным. Конференция стимулировала обмен опытом между национальными статистическими управлениями и другими важными заинтересованными сторонами и партнерами по вопросам налаживания партнерских отношений, доступа к данным и коммуникационной и агитационно-просветительской деятельности, а также по вопросу о необходимости учебной подготовки, повышения профессиональной квалификации и наращивания потенциала.

Совещание Глобальной рабочей группы по вопросам использования больших данных для подготовки официальной статистики. На своем первом совещании, состоявшемся 31 октября 2014 г., Глобальная рабочая группа по вопросам использования больших данных для подготов-

ки официальной статистики рассмотрела свой круг ведения и обсудила программу работы на 2015 г. В нынешний состав Глобальной рабочей группы входят 28 членов, представляющих как развитые, так и развивающиеся страны и различные международные и региональные организации.

Круг ведения и мандат Глобальной рабочей группы по вопросам использования больших данных для подготовки официальной статистики. В целом статистическое сообщество обязано изучать возможности использования новых источников данных для удовлетворения ожиданий общественности, связанных с получением более качественных конечных результатов и применением более совершенных и эффективных методов работы. Вместе с тем круг ведения и мандат Глобальной рабочей группы по вопросам использования больших данных для подготовки официальной статистики, несомненно, следует рассматривать также в контексте спроса на новые виды данных, которые необходимы для удовлетворения требований в области контроля за выполнением программы развития на период после 2015 г. и представления соответствующей отчетности.

Использование больших данных для подготовки официальной статистики полностью согласуется с основополагающими принципами официальной статистики. Например, принципы 1-й, 5-й и 6-й соответственно гласят: а) официальные статистические данные, которые удовлетворяют требованиям практической полезности, должны собираться и распространяться официальными статистическими учреждениями на беспристрастной основе в интересах реализации права граждан на общедоступную информацию; б) данные для статистических целей могут браться из всех типов источников, будь-то статистические обследования или данные административного учета, и статистические учреждения должны выбирать источник с учетом соображений качества, оперативности, издержек и бремени, ложащегося на респондентов; с) индивидуальные данные, собранные статистическими учреждениями для статистической об-

работки, независимо от того, касаются они физических или юридических лиц, должны носить строго конфиденциальный характер и использоваться исключительно в статистических целях. Исходя из этих принципов, статистические учреждения не только могут, но и практически обязаны изучать источники больших данных и добиваться их использования в статистических целях, если только при этом строго соблюдается неприкосновенность частной жизни и конфиденциальность. Большие данные могут в потенциале оказаться актуальными, своевременными и более затратноэффективными методами сбора данных по сравнению с традиционными методами.

В круге ведения Глобальной рабочей группы отражены выводы, сделанные в докладе Независимой консультативной группы экспертов по вопросу об информационной революции в интересах устойчивого развития (НКГЭ)², озаглавленном «A world that counts: mobilising the data revolution for sustainable development» (Мир, который считается: мобилизация информационной революции в интересах устойчивого развития) (ноябрь 2014 г.). В докладе подчеркивается, что статистическим управлениям необходимо будет осуществить преобразования, причем более быстрыми темпами, чем ранее, и продолжать процесс адаптации путем отказа от дорогостоящих и громоздких процессов подготовки данных, перехода к новым источникам данных, включая данные административного учета, поступающие из других государственных ведомств, и сосредоточения внимания на предоставлении таких данных, которые были бы пригодны для чтения человеком и машиной, совместимы с системами геопространственной информации и доступны с достаточной степенью оперативности для того, чтобы циклы подготовки данных совпадали с циклами принятия решений³.

В докладе Независимой консультативной группы экспертов также прямым текстом говорится о больших данных и в этой связи указывается на необходимость развития национального потенциала в области информатики в целях задействования возможностей больших данных и дополнения официальной

² Эта Независимая консультативная группа экспертов (НКГЭ) была назначена Генеральным секретарем 29 августа 2014 г. для подготовки доклада по вопросу об информационной революции в контексте программы развития на период после 2015 г. Консультативная группа опубликовала доклад 6 ноября 2014 г.

³ A world that counts, p. 9.

статистики. Для того чтобы информационная революция способствовала устойчивому развитию, необходимо выделение большего объема национальных ресурсов и оказание развивающимся странам международной поддержки. С опорой на двусторонние и международные механизмы координации и взаимодействия следует разработать и применять в растущих масштабах на гласной основе общие и стандартизированные виды использования больших данных для подготовки официальной статистики при всестороннем соблюдении применимых законов⁴.

Среди множества предложений, содержащихся в докладе Консультативной группы, выдвинуто также предложение о создании глобальной «объединенной сети информационных инновационных сетей» в целях сведения воедино усилий организаций и экспертов, работающих в этой сфере. Это способствовало бы применению передовой практики, позволяющей улучшить контроль за достижением новых целей в области устойчивого развития, выявлению областей, в которых с помощью единых информационных инфраструктур можно было бы решить проблемы недостаточного потенциала и повысить эффективность, развитию механизмов сотрудничества, выявлению серьезных пробелов в исследовательской работе и созданию стимулов, поощряющих новаторство.

В целом статистическим учреждениям следует выбирать источники данных с учетом соображений качества, оперативности, издержек и бремени, ложащегося на респондентов, и источники больших данных подпадают под эту категорию (как отмечалось выше). Большие данные могут в потенциале оказаться актуальными, своевременными и более затратно-эффективными методами сбора данных по сравнению с традиционными методами и могут обеспечить совпадение циклов подготовки данных с циклами принятия решений. Работа над вопросами использования больших данных должна способствовать применению передовой практики, позволяющей улучшить контроль за достижением новых целей в области устойчивого развития в рамках программы развития на период после 2015 г. Некоторые из новых показателей или их аппроксимативных индикаторов могли бы готовиться

на основе источников больших данных с повышенной степенью своевременности и мелкомасштабной социальной и геопространственной разбивкой.

Исходя из вышесказанного, Глобальная рабочая группа сформулировала свой круг ведения и мандат следующим образом: а) обеспечивать стратегическое видение, направленность и координацию глобальной программы использования больших данных для подготовки официальной статистики, в том числе показателей выполнения программы развития на период после 2015 г.; б) поощрять практическое использование источников больших данных, в том числе трансграничных данных, опираясь при этом на имеющиеся прецеденты и изыскивая пути для решения многочисленных проблем; в) содействовать наращиванию потенциала, проведению учебной подготовки и обмену опытом; г) налаживать коммуникационную и агитационно-просветительскую работу по вопросам применения больших данных в сфере политики, особенно для контроля за выполнением программы развития на период после 2015 г.; е) формировать доверие общественности в вопросах использования больших данных из частного сектора для подготовки официальной статистики.

Программа работы и ожидаемые конечные результаты. По итогам обсуждений, состоявшихся на Международной конференции и на совещании самой Глобальной рабочей группы, Глобальная рабочая группа договорилась о том, что ее программа работы будет состоять из нескольких направлений работы, курируемых конкретными целевыми группами, и что общую координацию этой работы будет осуществлять еще одна дополнительная группа. Во-первых, было решено продолжать работу в области использования трех классов источников больших данных, определенных на Международной конференции, а именно данных из сети мобильной телефонной связи, данных с изображений, полученных со спутников, и данных из социальных Интернет-сетей. Было достигнуто понимание о том, что при продвижении вперед работы в каждой из этих трех целевых групп будет поддерживаться взаимодействие как минимум с одним экспериментальным проектом.

⁴ A world that counts, p. 23.

Во-вторых, поскольку одной из постоянных тем на Международной конференции был доступ к данным и налаживание механизмов партнерства с частным сектором и другими сообществами, была создана целевая группа для более углубленного изучения этих взаимосвязанных тем. Эта группа займется также изучением возможности разработки всеобъемлющих соглашений о доступе к данным с поставщиками больших данных, действующими на мировой арене. Доступ к большим данным и их использование зависят также от наличия доверия среди общественности и распространения информации о преимуществах больших данных в целом и связанных с ними проблемах. Для этого была создана целевая группа по агитационно-просветительской и коммуникационной работе, которая займется также изучением стратегий сбора средств, с тем чтобы развивающиеся страны могли активно участвовать в осуществлении экспериментальных проектов.

В-третьих, в контексте программы развития на период после 2015 г. было дополнительно решено создать отдельную группу, которой будет специально поручено следить за взаимосвязями между показателями, необходимыми для контроля за достижением целей в области устойчивого развития, и видами использования больших данных. Наконец, было предложено и решено создать еще две группы: группу по учебной и профессиональной подготовке и наращиванию потенциала (ввиду того, что для обработки больших данных необходимы другие навыки по сравнению с навыками, которыми в настоящее время обладают работники статистических управлений); и группу по таким сквозным вопросам, как классификация и платформы. Последняя из указанных групп будет, в частности, заниматься разработкой классификации источников больших данных, опираясь при этом на классификацию, уже предложенную группой ЕЭК по большим данным. Общая координация работы этих восьми целевых групп будет осуществляться координационной группой, состоящей из руководителей всех групп.

Было указано, что этим целевым группам Глобальной рабочей группы следует учитывать опыт реализации проекта ЕЭК в области использования больших данных и материалы, которые готовятся целевыми группами ЕЭК

по большим данным, поскольку эти целевые группы ЕЭК завершат в ближайшее время свою работу, тогда как целевые группы Глобальной рабочей группы ее только начинают. Это дает возможность продолжить уже начатую работу по отдельным направлениям. Каждой целевой группе Глобальной рабочей группы необходимо будет выработать собственную программу работы, график и мероприятия в тесной консультации с координационной группой.

Ожидаемые конечные результаты деятельности Глобальной рабочей группы напрямую связаны с работой восьми целевых групп. Поэтому ожидается, что Глобальная рабочая группа подготовит ряд перечисленных ниже предложений и докладов и будет также участвовать в осуществлении ряда экспериментальных проектов. Глобальная рабочая группа продолжит также составление перечня проектов использования больших данных, возможно по линии своей координационной группы. Результаты первого проведенного обследования представлены в следующем разделе. Поэтому в 2015 г. ожидается достижение перечисленных ниже конкретных результатов:

- предложения и мероприятия в области учебной и профессиональной подготовки и наращивания потенциала в целях использования больших данных для подготовки официальной статистики;
- доклад о взаимосвязях между большими данными и целями в области устойчивого развития;
- стратегия агитационно-просветительской и коммуникационной деятельности, посвященной использованию больших данных для подготовки официальной статистики;
- доклад об использовании данных из сети мобильной телефонной связи для подготовки официальной статистики, в котором уточняются методологические, информационно-технологические проблемы и проблемы соблюдения неприкосновенности частной жизни, предлагаются пути их решения и перечисляются возможные виды использования этих данных для осуществления контроля за выполнением программы развития на период после 2015 г.;
- доклад об использовании данных из социальных Интернет-сетей для подготовки официальной статистики, в котором уточня-

ются методологические, информационно-технологические проблемы и проблемы соблюдения неприкосновенности частной жизни, предлагаются пути их решения и перечисляются возможные виды использования этих данных для осуществления контроля за выполнением программы развития на период после 2015 г.;

- доклад об использовании данных с изображений, полученных со спутников, и данных дистанционного зондирования для подготовки официальной статистики, в котором уточняются методологические, информационно-технологические проблемы и проблемы соблюдения неприкосновенности частной жизни, предлагаются пути их решения и перечисляются возможные виды использования этих данных для осуществления контроля за выполнением программы развития на период после 2015 г.;

- предложения по расширению доступа к источникам больших данных и налаживанию механизмов партнерства с частным сектором;

- доклад по сквозным вопросам: классификация, платформы и таксономия.

Помимо достижения конкретных конечных результатов, ожидаемых от различных целевых групп, Глобальная рабочая группа будет участвовать в ряде экспериментальных проектов в области использования больших данных для подготовки официальной статистики, а также продолжит работу над составлением перечня проектов использования больших данных, опираясь при этом на перечень проектов использования больших данных, составленный ЕЭК, и на результаты первоначального обследования, проведенного Статистическим отделом и ЕЭК, которые представлены в следующем разделе.

Результаты обследования стратегий использования больших данных и перечень проектов, составленный Статистическим отделом и ЕЭК. В сентябре 2014 г. Статистический отдел и ЕЭК провели совместное обследование проектов использования больших данных для подготовки официальной статистики. Цель состояла в проведении общего обзора действующих проектов использования больших данных в целях содействия более предметным

обсуждениям в рамках статистического сообщества в целом и в целях дальнейшей разработки программ в рамках Глобальной рабочей группы. Опросник был разослан всем учреждениям, участвующим в проекте ЕЭК в области использования больших данных, и членам Глобальной рабочей группы по вопросам использования больших данных для подготовки официальной статистики.

Проведение обследования преследовало две цели: сбор информации о завершенных, текущих или потенциальных проектах использования больших данных, в том числе информации о механизмах партнерства, источниках данных и инструментах, и сбор информации об организационных условиях осуществления таких проектов. Были распространены два разных опросника. Цель первого опросника состояла в сборе информации об общей организации, стратегиях и структурах управления в области использования больших данных, а второй касался фактических проектов использования больших данных. Для целей проведения этого обследования было сформулировано достаточно широкое определение больших данных, в соответствии с которым они были охарактеризованы как «источники данных огромного объема, скорости прироста и многообразия, требующие новых инструментов и методов сбора, обеспечения сохранности, систематизации и обработки на эффективной основе».

Опросник был разослан в общей сложности 78 национальным статистическим управлениям и 28 международным организациям⁵. На опросник по организационным условиям поступили ответы от 32 стран, и 24 страны ответили на опросник по индивидуальным проектам, представив информацию в общей сложности по 54 проектам. В дополнение к этому международные организации представили информацию по трем проектам. Пожалуй, можно с полным основанием утверждать, что на эти страны и международные организации приходится значительная доля завершенных и текущих проектов использования больших данных во всем мире. Тем не менее Глобальная рабочая группа рассматривает вопрос о проведении последующего глобального обследования для сбора информации о еще

⁵ Международным организациям было сообщено по электронной почте о формате обследования главным образом для того, чтобы уведомить их о факте проведения обследования.

большем числе проектов для составления перечня проектов использования больших данных. Полный аналитический доклад о выводах, сделанных по результатам этого обследования, представляется в виде справочного документа. Ниже вкратце обобщены некоторые из основных выводов.

Основные выводы: организационные условия. Первостепенное внимание в опроснике по организационным условиям уделялось общему опыту и механизмам обработки больших данных с точки зрения стратегии использования больших данных, руководящих и управленческих структур, оценки качества, вопросов соблюдения неприкосновенности частной жизни и конфиденциальности и нехватки специалистов. Полученные ответы подтверждают выводы, сделанные на Международной конференции по вопросам использования больших данных для подготовки официальной статистики, относительно главных проблем, стоящих перед национальными статистическими управлениями.

Результаты обследования показали, что хотя долгосрочная перспективная программа использования больших данных разработана лишь в ограниченном числе стран, в ряде стран в ближайшее время будет сформулирована стратегия в области использования больших данных. Кроме того, в некоторых странах созданы внутренние лаборатории, целевые или рабочие группы для осуществления экспериментальных проектов для определения того, могут ли большие данные использоваться как источник официальной статистики и если да, то какими способами. Однако в большинстве стран пока не определены рабочие процессы, которые позволили бы интегрировать источники и результаты использования больших данных в работу их статистической системы, и отсутствует четкая структура управления проектами использования больших данных.

Во многих странах предпринимаются шаги по налаживанию механизмов партнерства с целью изучения возможностей использования больших данных в рамках разработки стратегии в этой области. Для сведения к минимуму риска оказаться в стороне от процесса изучения этих новых технологий страны решили принимать активное участие в таких инициа-

тивах по международному сотрудничеству, как проект «Песочница» ЕЭК, Целевая группа Евростат по вопросам больших данных и Глобальная рабочая группа по вопросам использования больших данных для подготовки официальной статистики, созданная во исполнение решения Статистической комиссии.

Только в ограниченном числе стран действуют конкретные базовые положения по вопросам соблюдения неприкосновенности частной жизни при использовании больших данных, хотя большинство ответивших на опросник указали, что базовые положения по вопросам соблюдения неприкосновенности частной жизни, действующие в отношении традиционных видов статистики, применяются в настоящее время и к большим данным. Многие организации подчеркнули важность соблюдения принципов неприкосновенности частной жизни и конфиденциальности при использовании больших данных, даже с превышением формальных требований закона ввиду значимости этих принципов для общественности. Некоторые организации избегают проблемы соблюдения неприкосновенности частной жизни и конфиденциальности, производя все проверочные операции и манипуляции с данными в месте расположения поставщика больших данных, и затем им остается только препроводить полученные агрегированные величины национальным статистическим управлениям. Некоторые организации утверждают, что несмотря на необходимость введения строгих нормативных положений о защите данных, такие положения нередко служат препятствием к получению доступа к таким данным.

Хотя большинство респондентов признали наличие проблем, связанных с информационными технологиями, профессиональной подготовкой, законодательством и методологией, наибольшей проблемой, возникшей при осуществлении проектов использования больших данных, является узкий или ограниченный доступ к потенциальным наборам данных. Поскольку большие данные являются в большинстве случаев собственностью частного сектора (например, Интернет-компаний, операторов мобильной телефонной связи и банков), налаживание тесного взаимодействия с частным сектором приобретает крайне важное значение. Многие из участни-

ков этого процесса являются всемирными компаниями; следовательно, мировое статистическое сообщество могло бы использовать свою коллективную силу договаривающейся стороны для получения доступа к таким наборам данных.

Основные выводы: индивидуальные проекты. В опроснике по индивидуальным проектам содержались более подробные вопросы, относящиеся к потенциальным областям использования, статуса проекта, конечных результатов и извлеченного опыта, а также подробные вопросы, касающиеся механизмов партнерства, источников данных и используемых методов анализа данных, инструментов и навыков. Многие из проектов, по которым была собрана информация в ходе обследования, являются текущими проектами, предназначенными для проверочного и экспериментального использования больших данных. В

то же время отдельные проекты уже завершены, а другие - находятся на более или менее продвинутом этапе рассмотрения.

Что же касается потенциальных областей использования, то большинство проектов связаны с экономической и финансовой статистикой, демографической и социальной статистикой и статистикой цен (см. рисунок ниже). Стоит отметить, что многие проекты часто применимы к нескольким областям статистики. В рамках таких проектов один источник данных используется для разных областей статистики. Например, Национальный институт статистики и географии Мексики провел сбор выборочных сообщений, помещенных в сети «Twitter», и использовал полученные данные для различных статистических целей, а именно для подготовки показателей субъективного благополучия личности и статистики туризма и трансграничной мобильности.

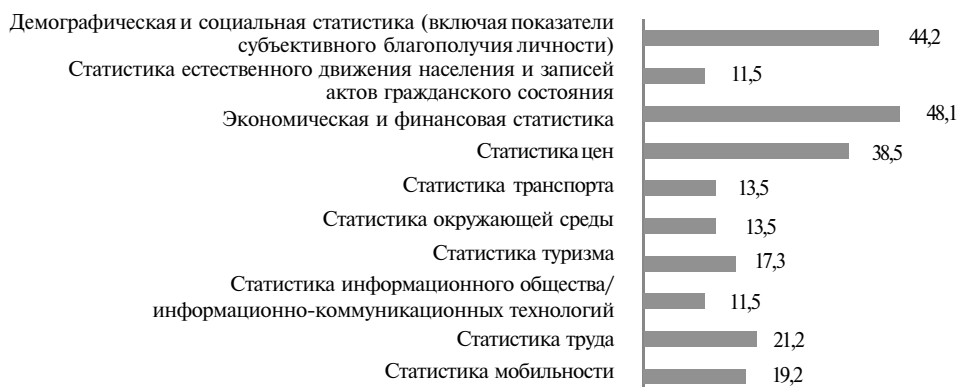


Рисунок. Потенциальные области использования в рамках данного проекта (в процентах)

Чуть более половины проектов осуществляются по линии механизмов партнерства. Наиболее частым партнером выступают коммерческие предприятия, второе место занимают прочие государственные учреждения. Наиболее типичным видом партнерства является партнерство с поставщиком данных, призванное обеспечивать доступ к данным. К этой категории относятся механизмы партнерства с компаниями мобильной телефонной связи, с поставщиками данных, полученных с помощью автоматического считывания показаний счетчиков, и с фирмами, занимающимися агрегированием данных из социальных Интернет-сетей. Например, для осу-

ществления проекта отслеживания картины суточной мобильности населения Национальный статистический институт Италии вступил в партнерство с поставщиком данных из сети мобильной телефонной связи. Федеральное статистическое управление Швейцарии вступило в партнерство с четырьмя крупными магазинами розничной торговли в этой стране для получения дважды в месяц данных сканирования товаров в целях расчета индексов потребительских цен на продовольственные и полупродовольственные товары.

Помимо установления механизмов партнерства с коммерческими поставщиками данных, в рамках многих проектов нацио-

нальные статистические управления наладили также партнерские отношения с аналитическими центрами. Такие отношения партнерства могут использоваться в целях приобретения конкретных технических знаний об источнике больших данных, преобразования необработанных данных в соответствии с параметрами заданной информационной структуры или применения методов анализа для получения определенных результатов. В то же время такой вид партнерства может использоваться и в целях предоставления людских ресурсов, способных выполнять трудоемкие задания. Например, Национальный институт статистики и географии Мексики вступил в партнерство с одним из мексиканских университетов для того, чтобы воспользоваться помощью большого числа студентов при классифицировании содержания сообщений, помещенных в сети «Twitter», в целях анализа оценок субъективного благополучия личности.

В подавляющем большинстве случаев участники обследования предпочитали использовать внутренние возможности для хостинга вместо приобретения внешних хостинг-услуг. В качестве главной причины использования внутреннего хостинга назывались соображения соблюдения неприкосновенности частной жизни. Внешний хостинг чаще использовался в тех случаях, когда данные уже считались достоянием общественности. С другой стороны, внутренний хостинг может стать фактором, ограничивающим выбор источника больших данных, поскольку средства, направляемые на обеспечение внутреннего хостинга, необходимо будет вкладывать в аппаратные средства, в программное обеспечение и в квалифицированных специалистов, и нехватка таких средств может оказаться серьезным препятствием для начала реализации проекта использования больших данных, особенно в условиях быстрых технологических изменений.

Заключение. Статистическое сообщество приступило к серьезному изучению возможностей, открывающихся в связи с применением источников больших данных для подготовки официальной статистики. Состоявшаяся в Пекине Международная конференция по вопросам использования больших данных для подготовки официальной статистики наглядно продемонстрировала возможности использования нескольких источников больших данных для подготовки официальной статистики в ряде разных областей и стимулировала обмен опытом между развитыми и развивающимися странами. Многие проблемы, стоящие перед национальными статистическими управлениями, имеют сходный характер, например получение доступа к данным, соображения, связанные с неприкосновенностью частной жизни, методологические вопросы, потребности в расширенных вычислительных ресурсах и отсутствие внутренних специалистов, которые занимались бы аналитикой больших данных, и указывают на необходимость глобального сотрудничества. Глобальная рабочая группа начала свою работу с определения методов своего функционирования и областей основного внимания. Она приступила также к процессу составления перечня текущих мероприятий, связанных с использованием больших данных для подготовки официальной статистики. По результатам обследования проектов использования больших данных был выявлен ряд интересных примеров, которые могут служить важным подспорьем для других организаций. Глобальной рабочей группе во взаимодействии с региональными рабочими группами необходимо разработать дальнейшие методические указания и осуществить ряд экспериментальных проектов в целях закрепления и ускорения прогресса в деле использования больших данных для подготовки официальной статистики.