Организация Объединенных Наций ECE/CES/2014/7

Экономический и Социальный Совет

Distr.: General
24 January 2014
Russian
Original: English

Европейская экономическая комиссия

Конференция европейских статистиков

Шестьдесят вторая пленарная сессия
Париж, 9−11 апреля 2014 года
Пункт 7 b) предварительной повестки дня
Итоги углубленных анализов, проведенных Бюро Конференции европейских статистиков

Углубленный анализ "больших данных"

Подготовлен временной Целевой группой по "большим данным" и секретариатом

Резюме

Настоящая записка является обновленным вариантом подробного аналитического документа по "большим данным". Углубленный анализ был проведен Бюро Конференции европейских статистиков на его совещании, состоявшемся в октябре 2013 года. Цель анализов, проводимых Бюро, заключается в повышении уровня координации статистической деятельности в регионе Европейской экономической комиссии Организации Объединенных Наций, выявлении пробелов или случаев дублирования усилий, а также рассмотрении новых вопросов.

Данная записка была подготовлена на основе итогов обсуждения, состоявшегося в ходе анализа. В ней содержится краткая информация о международной статистической деятельности, касающейся "больших данных", определяются вопросы и проблемы и приводятся рекомендации относительно того, каким образом международное статистическое сообщество могло бы решать их.

Информация об итогах этого анализа содержатся в документе ECE/CES/2014/7/Add.1

GE.14-20355 (R) 200314 240314

1320355

Просьба отправить на вторичную переработку

I. Резюме

1. Настоящий углубленный анализ посвящен реакции сообщества официальной статистики на явление "больших данных". Он дает определение "больших данных" и кратко описывает деятельность различных национальных и международных статистических организаций в этой области. В данном анализе в общих чертах излагаются выявленные на настоящий момент основные вопросы и проблемы и содержится вывод о том, что было бы более целесообразно, если бы международное статистическое сообщество решало эти вопросы на основе сотрудничества, не оставляя организации наедине с их проблемами. В заключительной части обзора сформулированы следующие три рекомендации:

a) конкретно определить основные приоритетные вопросы, касающиеся "больших данных", которые следует решать в рамках совместной деятельности международного статистического сообщества. Содержащееся в приложении проектное предложение отражает соображения, высказанные экспертами до настоящего времени в этой области. Было бы весьма желательно узнать мнение КЕС относительного того, охватывает ли это предложение надлежащие виды деятельности;

b) когда будут накоплены знания и опыт в области использования "больших данных", потребуется создать механизм для обмена этой информацией. Перечень связанных с "большими данными" видов деятельности, к составлению которого приступила неофициальная Целевая группа, следует консолидировать и расширять в качестве информационного ресурса для всего статистического сообщества;

c) два вышеуказанных вида деятельности должны находиться под контролем Группы высокого уровня по модернизации статистического производства и услуг с целью обеспечения достаточно стратегической направленности.

II. Введение

2. Бюро Конференции европейских статистиков (КЕС) регулярно проводит углубленный анализ избранных статистических областей. Цель анализа заключается в совершенствовании координации статистической деятельности в регионе ЕЭК ООН, выявлении пробелов или дублирования в работе, а также в рассмотрении новых вопросов. Анализ посвящен стратегическим вопросам и сосредоточен на проблемах статистических ведомств как концептуального, так и координационного характера. Настоящий документ служит основой для анализа; в нем содержится краткий обзор международной статистической деятельности в области "больших данных", указываются существующие вопросы и проблемы и содержатся рекомендации о возможных последующих мерах.

3. В нашем современном мире все больше данных генерируется в Сети и производится датчиками на все большем числе электронных устройств, окружающих нас. Объем данных и частота, с которой их производят, привели к появлению понятия "большие данные". Термин "большие данные" используется для описания наборов данных растущего объема, скорости разработки и разнообразия. Информационные источники, описываемые как "большие данные", часто являются в значительной степени неструктурированными, что означает, что они не следуют какой-то определенной заранее модели и/или не вписыва-

ются в традиционные реляционные базы данных. Помимо создания новых торговых возможностей в частном секторе, "большие данные" могут также представлять значительный интерес с точки зрения их использования в качестве исходных данных для официальной статистики: либо обособленно, либо в сочетании с более традиционными источниками данных, такими как выборочные обследования и административные регистры. Однако извлекать информацию из "больших данных" и использовать ее в процессе статистического производства нелегко.

4. "Большие данные" обладают потенциалом разработки более релевантной и своевременной статистики по сравнению с традиционными источниками официальной статистики. Официальная статистика практически полностью основывается на сборе данных в рамках обследований и получении административных данных от правительственных программ. Однако с "большими данными" дело обстоит иначе, поскольку большинство таких данных находится в открытом доступе или у частных компаний. В результате частный сектор может использовать эру "больших данных" для разработки все большего объема статистических данных, способных превзойти официальную статистику по своевременности и релевантности. Статистические организации вряд ли утратят свое фирменное наименование "поставщиков официальной статистики", однако если они останутся в стороне, то рискуют постепенно утратить свой авторитет и релевантность. Одним из больших преимуществ статистических организаций является наличие инфраструктуры для обеспечения точности, непротиворечивости и интерпретируемости производимых статистических данных. Задействуя соответствующие источники "больших данных" в своем процессе подготовки официальной статистики, статистические организации способны лучше других оценивать их точность, обеспечивать непротиворечивость всей системы официальной статистики и надлежащее толкование данных и при этом постоянно работать над их релевантностью и своевременностью. Это позволит сохранить роль и значимость официальной статистики.

5. Однако тема "больших данных" по-прежнему является относительно новой для многих статистических организаций, и существует неопределенность в отношении того, что они реально означают для официальной статистики и что является наиболее оптимальной реакцией на них.

III. Охват/определение рассматриваемой статистической области

6. Тема "больших данных" охватывает все виды статистической деятельности и может быть актуальной для всех статистических областей. С точки зрения Классификации статистической деятельности (пересмотренный вариант 1, октябрь 2009 года[1]) "большие данные", вероятно, наилучшим образом подходят для вида деятельности 4.3 (источники данных), однако их не так-то просто отнести к какой-либо более конкретной тематической области в этом разделе, если только не использовать весьма широкое определение "административных источников", в соответствии с которым "большие данные" можно было бы включить в тематическую область 4.3.5 (другие административные источники).

[1] http://www1.unece.org/stat/platform/display/disaarchive.

7. В рамках настоящего анализа "большие данные" рассматриваются в качестве источника информации, а не как выходные материалы статистических организаций, поскольку статистические материалы (пока) не отвечают критериям объема, скорости разработки и разнообразия, в связи с чем их действительно нельзя рассматривать в качестве "больших данных".

IV. Анализ международной статистической деятельности в данной области

A. Европейская экономическая комиссия Организации Объединенных Наций (ЕЭК ООН)

8. В октябре 2012 года ЕЭК ООН в сотрудничестве с Федеральной службой государственной статистики Российской Федерации (Росстат) организовала семинар высокого уровня на тему "Модернизация процесса производства статистических данных и услуг", который состоялся в Санкт-Петербурге[2]. После состоявшихся обсуждений, в ходе которых был затронут вопрос "больших данных", участники семинара сделали, в частности, следующий вывод:

"Использование "больших данных" становится все более насущной задачей. Официальному статистическому сообществу необходимо более глубоко понять соответствующие вопросы и разработать новые методы, средства и идеи для того, чтобы обеспечить эффективное использование источников "больших данных". Это предусматривает более тесную интеграцию с географическими данными и стандартами".

9. В качестве последующих мер Группе высокого уровня по модернизации статистического производства и услуг (ГВУ) было предложено подготовить "документ, поясняющий вопросы, связанные с использованием "больших данных" в сообществе официальной статистики". Для подготовки документа, посвященного этим вопросам, ГВУ при помощи ЕЭК ООН создало группу, в состав которой вошли ведущие международные эксперты. Итоговый документ под названием "Каково значение "больших данных" для официальной статистики?" был опубликован ГВУ в марте 2013 года[3] и спустя три месяца был представлен Конференции европейских статистиков на их пленарной сессии 2013 года.

10. Данный документ способствовал дальнейшему обсуждению вопроса об использовании "больших данных" в официальной статистике, в том числе на совместном совещании ЕЭК ООН/Евростата/ОЭСР/ООН-ЭСКАТО по управлению статистическими информационными системами (Париж и Бангкок, апрель 2013 года)[4]. Участники этого совещания указали, что использование "больших данных" является ключевым вопросом для официальной статистики, и отметили следующие основные аспекты:

- именно сейчас следует приступить к сотрудничеству по вопросу о "больших данных", поскольку организации, как правило, еще не создали соответствующих систем и могли бы разрабатывать их совместно;
- организации сталкиваются со схожими вопросами при использовании "больших данных", в связи с чем в интересах повышения эффективности

[2] http://www.unece.org/stats/documents/2012.10.hls.html.
[3] http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=77170614.
[4] http://www1.unece.org/stat/platform/display/msis/MSIS+2013.

соответствующей деятельности можно было бы объединить их усилия с целью нахождения общих решений. Это должно стать одним из приоритетов для ГВУ;

- статистические организации традиционно уделяют повышенное внимание подготовке согласованных динамических рядов, однако все более насущной становится также необходимость разработки кратковременных показателей для описания того или иного явления в стране в тот момент, когда оно возникает;
- опыт, накопленный в ходе использования административных источников, может оказаться полезным в отношении "больших данных";
- к "большим данным" следует применять многодисциплинарный подход, однако все группы в настоящее время рассматривают этот вопрос, исходя из своей собственной перспективы;
- в первоочередном порядке следует согласовать общую классификацию различных видов "больших данных";
- было бы полезно реализовать конкретный проект по подготовке конкретной статистики на основе "больших данных" и нахождению практических решений;
- следует создать виртуальную целевую группу с целью определения существующих вопросов и формулирования четкого проектного предложения, которое будет представлено ГВУ.

11. В мае 2013 года была создана временная целевая группа[5], которой было поручено выявить основные проблемы, связанные с использованием "больших данных" в целях официальной статистики, определить приоритетные меры и сформулировать проектное предложение. Эта группа осуществляла свою деятельность в мае и июне, действуя в виртуальном режиме с помощью вики-сайта и веб-конференций, и подготовила следующие материалы:

- проект предложения, предусматривающий три основные направления работы: изучение стратегических и методологических вопросов; анализ источников "больших данных" и международное тиражирование итоговых материалов с помощью компьютерной среды коллективного пользования; и проведение учебных и пропагандистско-просветительских мероприятий (см. приложение);
- проект классификации видов "больших данных"[6];
- спецификация перечня источников и проектов "больших данных" с опорой на вышеуказанную классификацию. Затем была начата работа по составлению этого перечня с наполнением его информацией о проектах, осуществляющихся в нескольких странах[7]; и которое будет продолжено в контексте деятельности ЕЭК ООН в области модернизации статистики.

12. После представления ежегодного доклада ГВУ на пленарной сессии КЕС 2013 года делегатам было предложено определить приоритеты для проектов ГУВ на 2014 год. Большинство выступавших признали необходимость осуществления международного сотрудничества в области упрощения использования "больших данных" в целях официальной статистики.

[5] http://www1.unece.org/stat/platform/display/msis/Members+of+the+task+team.
[6] http://www1.unece.org/stat/platform/display/msis/Classification+of+Types+of+Big+Data.
[7] http://www1.unece.org/stat/platform/display/msis/Big+Data+Inventory.

13. ГВУ обсудила и определила основные приоритеты на 2014 год вместе с представителями групп экспертов на ежегодном рабочем совещании по модернизации процесса статистического производства и услуг, которое состоялось 25−27 ноября 2013 года в Женеве. Большие данные были определены в качестве темы для крупного проекта международного сотрудничества на 2014 год, который будет осуществляться под контролем ГВУ.

B. Евростат

14. В настоящее время Евростат изучает потенциал использования "больших данных" в целях официальной статистики в таких областях, как статистика цен (подготавливаемая на основе данных о ценах, собираемых через Интернет) и статистика использования информационно-коммуникационных технологий (ИКТ).

15. Евростат активно участвует в работе Целевой группы, которой руководит ЕЭК ООН, а также в разработке окончательного проектного предложения. Сотрудники Евростата также подготовили для международных конференций ряд документов по этому вопросу.

16. Одно из заседаний ежегодного совещания ГДНСИ (Генеральных директоров национальных статистических институтов), состоявшегося в сентябре 2013 года в Гааге, Нидерланды[8], было посвящено вопросу "больших данных". В ходе этого заседания представители национальных статистических организаций и частных компаний выступили с соответствующими сообщениями. По итогам этого заседания был принят Схевингенский меморандум по "большим данным" и официальной статистике, в соответствии с которым членам Европейской статистической системы рекомендуется разработать стратегию использования "больших данных", обмениваться опытом и осуществлять сотрудничество на уровне Европейской статистической системы и за ее пределами. К середине 2014 года должны быть приняты план и "дорожная карта", которые будут учтены в программе работы Евростата.

17. В сотрудничестве с ЕЭК ООН Евростат провел рабочее совещание по вопросу "больших данных", которое состоялось в апреле 2014 года в Риме.

C. Организация экономического сотрудничества и развития (ОЭСР)

18. В настоящее время ОЭСР изучает возможность использования "больших данных" в таких областях, как показатели инноваций, качества интернет-связи и благосостояния/повышения качества жизни.

19. ОЭСР активно участвует в работе Целевой группы, которой руководит ЕЭК ООН, и в разработке окончательного проектного предложения. Сотрудники ОЭСР также подготовили для международных конференций документы по этому вопросу. Кроме того, ОЭСР опубликовала программный документ "Изучение основанных на данных инноваций как нового источника роста: обзор стратегических вопросов, связанных с "большими данными"[9].

[8] http://www.cbs-events.nl/dgins2013/.

[9] http://www.oecd-ilibrary.org/science-and-technology/exploring-data-driven-innovation-as-a-new-source-of-growth_5k47zw3fcp43-en.

D. Статистический отдел Организации Объединенных Наций

20. В ходе сессии Статистической комиссии 2013 года Статистический отдел Организации Объединенных Наций организовал однодневное параллельное мероприятие под названием «Роль "больших данных" в формировании политики, развитии и подготовке официальной статистики»[10]. В рамках этого мероприятия представители национальных статистических учреждений, международных организаций и частных компаний выступили с соответствующими сообщениями.

21. В повестку дня сессии Статистической комиссии 2014 года был также включен такой вопрос, как «"Большие данные" и модернизация статистических систем»[11], по итогам обсуждения которого было предложено "создать глобальную рабочую группу по вопросам использования "больших данных" в целях подготовки официальной статистики, деятельность которой дополняла бы работу, проводимую региональными комиссиями, и которая занималась бы решением вопросов, имеющих общемировую значимость.

E. Всемирный банк

22. Всемирный банк организовал ряд мероприятий по вопросу "больших данных", включая проведенное в октябре 2012 года мероприятие на тему "Преобразование "больших данных" в мощное влияние"[12] и состоявшуюся в декабре 2012 года прямую интернет-трансляцию обсуждения на тему "Что произойдет тогда, когда "большие данные" будут удовлетворять требованиям официальной статистики?"[13].

F. Другие мероприятия

23. За рамками официальной статистики проводятся многочисленные мероприятия и обсуждения по вопросу "больших данных". В настоящее время стремительно увеличивается число консультантов по "большим данным", а также средств программного обеспечения, конкретно предназначенных для работы с "большими данными". Совершенно ясно, что в ходе разработки ответных мер на появление "больших данных" международному официальному статистическому сообществу следует активно следить за внешними разработками и определять, каким образом они могли бы использоваться в нашей деятельности.

V. Практика стран

24. Целевая группа по "большим данным" выявила широкий спектр мероприятий в области использования "больших данных", проводимых в странах-участницах. Большинство этих мероприятий находится на стадии планирования и экспериментирования с целью определения возможностей использования источников больших данных в целях статистического производства. Однако небольшое число стран приступает к реализации следующего этапа и движется

[10] http://unstats.un.org/unsd/statcom/statcom_2013/seminars/Big_Data/default.html.
[11] http://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf.
[12] http://www.linkedin.com/groups/World-Bank-Event-Turning-Big-137043.S.177763767.
[13] http://live.worldbank.org/what-happens-when-big-data-meets-official-statistics-live-webcast.

в направлении регулярной подготовки статистической информации на основе "больших данных".

25. Одна из выявленных проблем заключается в отсутствии механизма для обмена информацией о текущей и планируемой деятельности. В этой связи было предложено разработать перечень проектов и ресурсов в области "больших данных".

VI. Влияние кризиса на статистику

26. Финансовый кризис, начавшийся в 2009 году, послужил мощным стимулом для статистических организаций к началу поиска путей повышения эффективности своей деятельности и сокращения расходов, связанных с подготовкой данных. Сбор данных традиционно является одним из наиболее дорогостоящих элементов процесса подготовки статистической информации, что обуславливает повышение интереса к альтернативным источникам данных, включая "большие данные".

VII. Вопросы и проблемы

27. В документе ГВУ "Каково значение "больших данных" для официальной статистики?" были выявлены нижеследующие вопросы и проблемы.

A. Трудности законодательного порядка

28. Законодательство некоторых стран может предоставлять право доступа к данным, содержащимся как в государственных, так и негосударственных источниках, в то время как в других странах право доступа предоставляется только к данным, имеющимся в распоряжении у государственных органов. Это может приводить к ограничениям доступа к некоторым видам "больших данных".

29. Даже если законодательство предусматривает доступ ко всем видам данных, в различных странах может понадобиться в разной степени обосновывать статистическую цель получения доступа к этим данным.

B. Связанные с конфиденциальностью трудности

30. Конфиденциальность обычно определяется как право людей контролировать или регулировать содержание касающейся их информации, которая может быть предана гласности. Право на конфиденциальность является одной из основ демократии. Проблема, связанная с "большими данными", заключается в том, что пользователи услуг и устройств, генерирующих данные, скорее всего не знают, что они это делают и/или для каких целей эти данные могут использоваться. Увязка данных приведет к росту их объема, а также озабоченности по поводу защиты конфиденциальности.

C. Финансовые трудности

31. Как представляется, статистическим организациям придется платить за приобретение "больших данных", особенно "больших данных", держателем которых является частный сектор, в частности если законодательство обходит

молчанием финансовые условия приобретения внешних данных. Поэтому статистическим организациям следует принимать взвешенные решения, учитывающие, с одной стороны, требования качества (которые включают в себя такие аспекты, как релевантность, своевременность, точность, целостность, доступность и интерпретируемость данных) и, с другой стороны, издержки и сокращение нагрузки на респондентов. Издержки могут быть существенными, однако потенциальные выгоды могут значительно перевешивать их, поскольку из "больших данных" можно извлекать информацию, которая способна повысить эффективность государственных программ (например, системы здравоохранения). В этом случае могут применяться правила государственных закупок.

D. Управленческие трудности

32. Использование "больших данных" в целях официальной статистики может означать поступление в статистические организации все большего объема информации, которая должна отвечать требованиям политики и директив в области администрирования и защиты информации.

33. Еще одна управленческая трудность связана с людскими ресурсами. Эксперты по данным применительно к "большим данным", появляющимся в частном секторе, по всей видимости, пока еще не установили связей с сообществом официальной статистики. Не исключено, что статистическим организациям придется проводить внутренние и национальные опросы (научных кругов, государственного и частного секторов) для выявления "ученых по данным" и подключения их к сфере официальной статистики.

E. Методологические трудности

34. Основной проблемой, связанной с "большими данными", является их репрезентативность. Трудность определения целевой совокупности, обследуемой совокупности и основы выборки ставит под угрозу традиционное представление официальных статистиков о целевой (и конечной) совокупности и их статистический анализ. При традиционном обследовании статистики определяют целевую/обследуемую совокупность, создают основу выборки для охвата этой совокупности, формируют выборку, собирают данные и т.д. Затем они строят таблицу и заполняют ее данными в строго структурированной форме. В случае "больших данных" сначала появляются сами данные, на основе которых официальным статистикам предстоит построить таблицу. В связи с этим возникает следующий вопрос: неужели это единственный способ построения целостной и комплексной системы национальной статистики? Возможно, пришло время выйти за рамки таблицы?

35. Еще одна проблема носит технологический и методологический характер. При анализе все большего объема данных традиционные статистические методы, разработанные для крайне тщательных анализов небольших выборок, становятся неэффективными. В самых простых случаях они оказываются просто недостаточно быстрыми. Требуется разработать новые методы и инструменты, например такие, как:

a) методы быстрого извлечения информации из массивов имеющихся данных, такие как методы визуализации и методы извлечения данных, текста и потокового извлечения знаний, которые способны «ужимать "большие дан-

ные"». Увеличение вычислительной мощности компьютеров является первоначальной мерой поддержки на этом этапе;

b) методы, способные интегрировать информацию, обнаруженную в статистическом процессе, такие как увязка в массовом масштабе, интеграция данных, а также статистические методы, конкретно подходящие для больших наборов данных. Необходимо разработать методы, которые позволили бы быстро получать надежные результаты при работе с очень большими наборами данных.

36. Использование "больших данных" в официальной статистике, безусловно, порождает потребность в новых технических приемах. Эти технические приемы связаны с нижеследующими методологическими проблемами, требующими решения:

a) показатели качества материалов, полученных из внешних данных, с трудом поддающихся регулированию. Зависимость от внешних источников ограничивает набор показателей, которые могут указываться при сравнении с результатами, полученными при использовании целевых методов сбора информации;

b) ограниченная сфера применения и ценность данных, полученных из внешних источников;

c) сложности, связанные с интеграцией информации из различных источников для получения высококачественных продуктов;

d) трудность определения ожидаемой полезности в отсутствие замкнутой обратной связи, существующей в коммерческих организациях.

F. Технологические трудности

37. Необходимо разработать новые инструменты для подключения напрямую приложений сбора данных и обработки данных с административными данными. Сбор данных в реальном масштабе времени и режиме, близком к реальному, способен максимально повысить потенциал данных за счет создания новых возможностей использования данных, поступающих из таких источников высокоскоростных данных, как:

a) коммерческие данные (операции по кредитным картам, онлайновые операции, продажи и т.д.);

b) отслеживающие устройства (сотовые телефоны, системы глобального позиционирования, камеры наблюдения, приложения) и физические датчики (метеорологические и энергетические датчики, датчики движения и загрязнения окружающей среды и т.д.);

c) социальные сети ("Твиттер", "Фейсбук" и т.д.) и поисковые системы (поиск в Интернете, просмотр веб-страниц);

d) общинные данные (сообщения граждан или краудсорсинговые данные).

38. В эпоху "больших данных" такое изменение парадигмы сбора данных позволяет собирать и интегрировать множественные виды данных из многочисленных различных источников. Комбинирование традиционных источников данных, таких как обследования и административные данные, с "большими данными" может создавать новые проблемы и открывать новые возможности.

VIII. Выводы и рекомендации

39. Совершенно ясно, что официальное статистическое сообщество только приступает к изучению потенциальных проблем и преимуществ, связанных с "большими данными". Если каждая организация будет заниматься этим обособленно, это приведет к снижению эффективности статистической системы на глобальном уровне. Поэтому по итогам настоящего углубленного анализа можно сформулировать следующие основные рекомендации:

a) необходимо конкретно определить основные приоритетные вопросы, касающиеся "больших данных", которые следует решать в рамках совместной деятельности международного статистического сообщества. Проект ГВУ по "большим данным" направлен на решение этой задачи;

b) когда будут накоплены знания и опыт, касающиеся использования больших данных, необходимо будет создать механизм для обмена этой информацией. Перечень связанных с "большими данными" мероприятий, к составлению которого приступила неофициальная целевая группа, подлежит консолидации и расширению в качестве информационного ресурса для всего статистического сообщества;

c) два вышеуказанных вида деятельности должны находиться под контролем Группы высокого уровня по модернизации статистического производства и услуг с целью обеспечения достаточно стратегической направленности.