Организация Объединенных Наций ECE/CES/2017/8

Экономический и Социальный Совет

Distr.: General
25 March 2017
Russian
Original: English

Европейская экономическая комиссия

Конференция европейских статистиков

Шестьдесят пятая пленарная сессия
Женева, 19–21 июня 2017 года
Пункт 8 a) предварительной повестки дня
Итоги углубленных анализов, проведенных Бюро Конференции европейских статистиков

Углубленный анализ темы «Интеграция данных»

Записка Группы высокого уровня по модернизации официальной статистики

Резюме

Настоящий документ послужил основой для углубленного анализа темы «Интеграция данных», проведенного Бюро Конференции европейских статистиков в феврале 2017 года.

В данном документе отражен опыт, приобретенный в ходе реализации проекта по интеграции данных в 2016 году, осуществляемого под эгидой Группы высокого уровня по модернизации официальной статистики Европейской экономической комиссии Организации Объединенных Наций. Он содержит общий обзор экспериментов по интеграции данных и описывает возможности, проблемы и вызовы в этой области.

В разделе VI приводятся выводы и рекомендации. Информация об итогах этого анализа содержится в документе ECE/CES/ 2017/8/Add.1.

GE.17-04862 (R) 280417 010517

1704862

I. Резюме

1. В настоящем анализе описывается опыт, накопленный в ходе экспериментов по интеграции данных, проведенных в рамках проекта по интеграции данных в 2016 году, в котором принимали участие национальные статистические управления Бразилии, Венгрии, Италии, Канады, Колумбии, Нидерландов, Новой Зеландии, Польши, Словении и Сербии. Были изучены различные типы интеграции данных: результатов обследований и административных источников, результатов обследований и новых источников данных (в том числе больших данных), традиционных источников с геопространственной информацией и интеграции данных для проверки достоверности официальных статистических данных. В контексте этих типов интеграции был проведен ряд экспериментов, позволивших выявить возможности, проблемы и вызовы и сформулировать многочисленные рекомендации и извлечь различные уроки.

2. К числу многочисленных возможностей, открываемых интеграцией данных, относится получение новых или более своевременных статистических данных, более дешевых в разработке и позволяющих снизить нагрузку на респондентов и потенциально более высокого качества. К числу основных выявленных проблем относятся: стабильность источников данных, необходимые компетенции и технологии, концептуальные различия, потребности в метаданных, управление качеством, эффективное партнерство, переход от тестирования к производству, расходы по реализации и эксплуатационные расходы, избежание дублирования и отношение общественности к проблеме конфиденциальности. Появление все большего числа источников данных и рост мощностей ИТ информационных технологий и инфраструктуры данных усиливают необходимость интеграции различных источников.

3. Интеграция может охватывать многочисленные и разнообразные типы данных и различные источники. Поэтому не существует никаких единых рекомендуемых методов для интеграции всех видов данных. Однако имеется несколько различных стандартных процессов интеграции данных. Проекты по интеграции данных могут предусматривать схожую последовательность шагов, таких как выявление потребностей и уточнение бизнес-требований, определение партнеров, отбор потенциальных источников данных, изучение методологий и качества, анализ издержек, выгод и рисков и получение данных и необходимых инструментов, кадров и ресурсов. После осуществления экспериментов их результаты должны оцениваться, а методы и подходы – совершенствоваться для превращения их в тиражируемое производственное решение.

4. Разработка рамок качества для интегрированных данных и удовлетворение потребностей в метаданных имеют важное значение для обеспечения обещанной официальной статистикой полезности, получаемой в результате интеграции источников данных. По своей природе, интеграция данных требует налаживания различных типов партнерства, что создает дополнительные вызовы для НСУ. Важное значение имеет развитие эффективной системы управления проектами по интеграции данных. В дополнение к основным статистическим компетенциям для разрешения и согласования и использования внешних источников данных, а также информирования о нем необходимы новые компетенции в сфере ИТ, компьютерные приложения и аппаратные средств и другие знания.

5. Реализация различных экспериментов по интеграции данных привела к разработке большого числа рекомендаций, касающихся, например:

- заручения поддержкой руководства высокого уровня и информирования руководства статистического управления о проекте и его целях и результатах;
- сохранения четкости целей и разъяснения важности проекта, спецификации характеристик проблемы, которую предстоит решить;

- налаживания эффективного сотрудничества с партнерами/поставщиками данных: четкое определение потребностей в данных, учет общих целей учреждений (сбор данных только один раз, сокращение излишних расходов), заключение соглашений;
- сотрудничества с пользователями данных;
- проведения консультаций с другими экспертами, поиска дополнительной информации о возможных методах и решениях для практического применения; обмена сотрудниками для приобретения опыта и изучения передовой практики;
- учета международных рекомендаций и стандартов в области статистики;
- обмена результатами;
- начала работы с выборки данных, изучения данных до начала эксперимента, сохранения терпения и настойчивости, рассмотрения широкого спектра решений; и
- измерения качества данных.

6. Реализация проекта по интеграции данных Группы высокого уровня по модернизации официальной статистики (ГВУ-МОС) продолжается в 2017 году.

II. Введение

7. Бюро Конференции европейских статистиков (КЕС) регулярно проводит углубленный анализ избранных областей статистики. Цель такого 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.

8. Бюро КЕС отобрало тему «Интеграция данных» в октябре 2015 года для углубленного анализа на своем совещании в феврале 2017 года. Настоящий документ обеспечивает основу для анализа; в нем содержится краткий обзор международной статистической деятельности в избранной области, указываются существующие вопросы и проблемы и содержатся рекомендации в отношении возможных последующих мер.

9. По сравнению с традиционными подходами интеграция данных позволяет производить статистические данные более своевременно и с более высокой степенью детализации и регулярностью. В 2015 году Группа высокого уровня по модернизации официальной статистики (ГВУ-МОС) признала, что перед организациями официальной статистики встала задача создания потенциала, необходимого для интеграции новых источников данных в их процессы статистического производства. В результате был разработан проект по интеграции данных ГВУ-МОС 2016 года. Настоящий анализ опирается на опыт и уроки, извлеченные странами – участницами проекта.

10. В разделе III настоящего документа описываются типы интеграции данных, а в разделе IV – общие рамки для интеграции данных и широкий подход к реализации проектов по интеграции данных. В разделе V кратко описываются проблемы, возможности, вызовы, меры профилактики рисков, рекомендации и необходимые компетенции и ресурсы. В последнем разделе содержатся выводы и краткие рекомендации, а также предложение для Бюро КЕС. Краткая информация о каждом предложенном эксперименте размещена на вики-сайте проекта по интеграции данных ГВУ-МОС 2016 года по адресу www1.unece.org/stat/platform/x/HQazBw.

III. Тематический охват/определение рассматриваемой статистической области

11. Цель проекта интеграции данных ГВУ-МОС 2016 года заключалась в приобретении опыта, который бы позволил разработать общие рекомендации и руководящие указания по интеграции данных и соответствующим рамкам качества. Проект предусматривал объединение ресурсов для проведения совместных практических мероприятий.

12. Существует множество возможных типов интеграции данных с использованием многочисленных комбинаций источников данных и подходов к моделированию. Охватить все типы интеграции данных в рамках одногодичного проекта было бы вряд ли осуществимо. Поэтому было принято решение сосредоточить внимание на экспериментах, охватывающих четыре основных типа интеграции данных, описываемых ниже (в скобках указано название страны, которая предложила проект и участвовала в нем, однако участие в реализации эксперимента могли принимать и другие страны).

13. Были отобраны следующие типы интеграции и осуществлены следующие эксперименты:

1. Интеграция результатов обследований и административных источников

a) Интеграция информации об образовании на основе географического местоположения школ (Колумбия);

b) интеграция потенциальных источников информации для подготовки статистических данных о вакантных рабочих местах (Венгрия);

c) увязка статистического регистра занятости и обследования рабочей силы (Словения).

2. Интеграция новых источников данных, таких как большие данные, и традиционных источников

a) Стратегии веб-сбора данных в целях официальной статистики – тематическое исследование Новой Зеландии и обзор опыта других стран (Новая Зеландия);

b) интеграция полученных с веб-сайтов данных для составления статистики цен (Венгрия);

c) интеграция данных сканирования и полученных с веб-сайтов данных о ценах для обеспечения сопоставимого на международном уровне измерения цен (Новая Зеландия);

d) расчет и сопоставление индексов цен на основе различных источников больших данных в разных странах (Новая Зеландия);

e) интеграция потенциальных источников информации для подготовки статистических данных о вакантных рабочих местах (Венгрия).

3. Интеграция геопространственной и статистической информации

a) Интеграция пространственных объектов, используемых в статистике и геодезии «Модель уровня 10», для согласования систем пространственной привязки статистики и геодезии (Польша);

b) анализ существующих моделей интеграции геопространственной и статистической информации (Колумбия).

4. Проверка достоверности официальной статистики с использованием данных из других источников

a) Сравнительный анализ данных о доходах, полученных в рамках обследования доходов Новой Зеландии, и административных данных (Новая Зеландия);

b) увязка статистического регистра занятости и обследования рабочей силы (Словения).

14. В рамках проекта по интеграции данных ГВУ-МОС 2016 года выделялся и пятый тип интеграции данных, «Интеграция микро- и макроданных», но эксперименты в этой области не проводились.

15. Некоторые мероприятия охватывали несколько типов интеграции данных. Проводились также межсекторальные мероприятия, направленные на получение и подготовку наборов данных и обобщение опыта, касающегося партнерства, методологий, рамок качества и потребностей в метаданных. Кроме того, сквозным направлением деятельности в рамках проекта являлось выявление и тестирование инструментов, использовавшихся для интеграции данных в различных экспериментах, перечисленных выше. Диаграмма 1 иллюстрирует общую структуру проекта.

Диаграмма 1
Структура проекта по интеграции данных ГВУ-МОС 2016 года

16. Настоящий анализ также основан на информации, полученной в рамках проекта по интеграции данных ГВУ-МОС 2016 года, и опыте, накопленном в результате реализации других проектов по интеграции данных, уже завершенных или находящихся в стадии осуществления. В нем рассматриваются вопросы и извлеченные уроки с целью разработки руководящих указаний в целях оказания поддержки и обеспечения продвижения интеграции данных в области официальной статистики. В настоящем анализе освещаются некоторые вопросы, которые, возможно, полезно было бы принимать во внимание в целях реализации успешных проектов по интеграции данных.

17. Свой вклад в реализацию проекта внесли следующие статистические учреждения: Бразильский институт географии и статистики (БИГС), Центральное статистическое управление (ЦСУ) Польши, Центральное статистическое управление Венгрии (ЦСУВ), Итальянский национальный институт статистики (ИСТАТ), Национальный административный департамент статистики (НАДС) Колумбии, Статистическое управление Республики Сербия (СУРС), Статистическое управление Словении (СУС), Статистическое управление Канады, Статистическое управление Нидерландов (СУН), Евростат и Европейская экономическая комиссия Организации Объединенных Наций (ЕЭК ООН).

IV. Обзор экспериментов по интеграции данных

18. В настоящем разделе содержатся краткий обзор рамок проектов по интеграции данных и описание проектов, которые были осуществлены в отношении различных типов интеграции. Краткая информация о каждом предложенном эксперименте размещена на вики-сайте проекта интеграции данных ГВУ-МОС 2016 года по адресу www1.unece.org/stat/platform/x/HQazBw.

19. Практическая работа по реализации экспериментов начинается с данных. В отсутствие надлежащих наборов данных, которые могли бы совместно использоваться организациями, могут быть получены или разработаны выборочные или синтетические наборы данных для обмена методами, инструментами и опытом интеграции данных. Для обработки и проверки источников массовых (больших) данных была предоставлена тестовая зона Центра высокопроизводительных вычислений Ирландии. Речь идет об относительно открытой среде, которая не предназначена для работы с персональными и иными конфиденциальными данными. При интеграции данных должны учитываться различные статистические, методологические, правовые и этические вопросы, однако использование наборов данных, предназначенных только для экспериментов, облегчает эту задачу. По мере совместной разработки соответствующих методов, процессов и инструментов можно будет перейти к защищенным средам индивидуальных организаций для проведения дальнейшего тестирования с использованием реальных данных.

A. Рамки для интеграции данных

20. Существует множество возможных типов интеграции данных, и для каждого из них – многочисленные комбинации источников данных и подходов к моделированию. Интеграция данных может проводиться на микроуровне, на уровне общего знаменателя и на агрегированных (макро-) уровнях с помощью методов моделирования или их определенного сочетания.

21. Выделяется пять общих типов интеграции:

a) административных источников с результатами обследований и другими традиционными данными;

b) новых источников данных (таких, как «большие данные») с традиционными источниками данных;

c) геопространственных данных со статистической информацией;

d) данных микроуровня с данными макроуровня; и

e) для проверки достоверности официальной статистики с использованием данных из других источников.

22. Новые возможности, открываемые появлением многочисленных источников данных, и проблемы с использованием обследований в качестве главного подхода к разработке официальной статистики привели директора Американского комитета по национальной статистике, Конни Сайтро, к следующему выводу: «Мы должны перейти от парадигмы разработки наилучших по возможности оценок на основе обследований к разработке наилучших по возможности оценок для удовлетворения потребностей пользователей с использованием множественных источников данных» (Citro, 2014)[1]. Таким образом, задача заключается в интеграции различных наборов несогласованных данных и производстве стабильных продуктов с использованием зачастую нестабильных, постоянно меняющихся исходных данных. Вместо того чтобы пытаться готовить наилучшую по возможности статистику с помощью отдельного обследования,

[1] (2014) Citro, C F. From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, 40(2), 137-161.

официальной статистике следует стремиться к нахождению наилучшего сочетания источников для разработки показателей/статистики, которые наилучшим образом будут удовлетворять потребностям пользователей.

23. Необходимо изучить возможности использования результатов обследований, административных данных, «больших данных» и других нетрадиционных источников. Несмотря на ряд предпринимавшихся попыток интеграции данных из различных источников для подготовки статистики, универсальная методология или рамки качества еще не созданы. В целях решения этой неотложной и сложной общей задачи проект ГВУ-МОС был направлен на объединение ресурсов для начала применения более системного подхода к разработке новых общих рамок статистического производства.

24. Реализация проектов по интеграции данных требует решения многочисленных задач, таких как определение потребностей, налаживание партнерств, получение данных; поиск надлежащих подходов к моделированию и управление качеством, рисками, сопоставимостью и определение потребностей в метаданных. В рамках проекта для организации вопросов, требующих рассмотрения, был определен набор широких тем. Таковыми являются:

- бизнес-требования;
- возможности;
- проблемы;
- смягчение рисков;
- стандартные процессы;
- рекомендуемые методы;
- соображения ИКТ;
- качество;
- стандарты;
- потребности в метаданных;
- соответствующая работа в рамках других проектов/организаций;
- компетенции;
- ресурсы;
- партнерства;
- управление;
- продвижение и информационная работа; и
- рекомендации.

25. Участники экспериментов изучили эти темы, обобщили извлеченные уроки и, в соответствующих случаях, разработали рекомендации. Данный проект направлен на обобщение опыта с целью выработки рекомендаций более общего характера, касающихся более одного типа интеграции данных.

B. Интеграция результатов обследований и административных источников

26. Некоторые страны уже накопили богатый опыт интеграции результатов обследований и административных источников данных. В этой области были реализованы совместные проекты, например, в Евростате.

27. Административные данные могут быть доступны уже в течение некоторого времени, но не использоваться. Они могут быть интегрированы с использованием увязки записей или статистического согласования или методов моде-

лирования. Это может предусматривать объединение или комбинирование результатов различных обследований, включая обследования, не проводимые самим НСУ.

28. При данном типе интеграции возникают определенные общие проблемы. Качество набора административных данных может быть достаточно хорошим для административных целей, но недостаточным для статистических. Преобразование наборов административных данных в наборы статистических данных может потребовать повышения качества и устранения концептуальных различий, особенно в тех случаях, когда административные данные планируется использовать напрямую. В случае обследований, проводимых с использованием данных из административных источников, чрезвычайно важно свести все данные (результаты обследований и данные административных источников) в одну базу данных.

29. Примеры источников, которые потенциально могут стать объектом интеграции, включают в себя: обследования рабочей силы и регистр социального страхования и/или регистры образования, данные министерства культуры и культурных ассоциаций для подготовки статистики посещаемости музеев. Существует ряд примеров административных данных, комбинируемых с результатами обследований для разработки показателей, данные по которым традиционно собираются в рамках переписей населения.

1. Краткое описание экспериментов

30. Уже были проведены следующие эксперименты:

- интеграция потенциальных источников информации для подготовки статистических данных о вакантных рабочих местах;
- увязка статистического регистра занятости и обследования рабочей силы; и
- система просмотра информации и географического местоположения школ.

31. Эксперименты, связанные с этим типом интеграции данных, как правило, предусматривали следующие шаги:

- определение статистических рамок и границ работы;
- отбор административных источников и набора результатов обследований;
- определение ситуации использования данных из административных данных вместо результатов обследования (меньшая совокупность) или добавления дополнительных переменных из административного обследования (сокращенный вопросник) или использования и того и другого;
- определение планируемых продуктов интеграции;
- разработка и тестирование процесса преобразования административных данных в статистические данные. Должен быть разработан метод очистки;
- разработка и тестирование процесса интеграции административных данных и результатов обследований; и
- оценка статистических продуктов, полученных в результате интеграции наборов административных данных и наборов результатов обследований.

2. Соответствующая работа в рамках других проектов/организаций

32. Интеграция результатов обследований и административных источников не является новой темой для официальной статистики. Работа в этой области ведется уже на протяжении многих лет различными организаций, в том числе в рамках программы работы КЕС. Из числа текущих проектов можно упомянуть следующие:

- база знаний ЕЭК ООН ASSIST по использованию административных и вторичных источников в целях статистики www1.unece.org/stat/platform/display/adso/ASSIST;
- Проект ESS.VIP ADMIN: цели проекта включают в себя оказание поддержки государствам-членам ЕС в извлечении выгод (сокращение расходов и нагрузки, увеличение объема доступных данных), связанных с использованием административных источников данных в целях подготовки официальной статистики, и гарантирование качества продуктов, производимых с использованием административных источников, в частности сопоставимости статистических данных, необходимых для целей Европейского союза. https://ec.europa.eu/eurostat/cros/content/essvip-admin-administrative-data-sources_en;
- Проект ESSnet по интеграции данных: основное внимание в рамках проекта уделялось методологиям интеграции данных (увязка записей, статистическое согласование, процессы микроинтеграции) и статистическим аспектам, которые необходимо учитывать для обеспечения применения этих методов конкретно НСИ. http://ec.europa.eu/eurostat/cros/content/data-integration-finished_en;
- Проект ESSnet: интеграция результатов обследований и административных данных. Цель проекта состояла в поощрении распространения знаний и применения на практике эффективных методов совместного использования существующих источников данных при подготовке официальной статистики. http://ec.europa.eu/eurostat/cros/content/isad-finished_en;
- Евростат (2013): использование регистров в контексте ОДУЖ-ЕС: вызовы и возможности. http://ec.europa.eu/eurostat/documents/3888793/5856365/KS-TC-13-004-EN.PDF;
- ЕЭК ООН (2007): Статистика на основе регистров в Североевропейских странах. Обзор передовой практики с уделением особого внимания демографической и социальной статистике. http://unstats.un.org/unsd/dnss/docViewer.aspx?docID=2764;
- Проект ЕЭК ООН по показателям качества для Типовой модели производства статистической информации (ТМПСИ). Речь идет о текущем проекте, направленном на разработку показателей качества для мониторинга качества процесса статистического производства для каждого из этапов ТМПСИ, включая подэтап «интеграция данных». В рамках этого проекта в настоящее время осуществляются обзор и обновление показателей качества с целью включения использования административных данных в процесс производства официальной статистики. Информация о текущей работе размещена по адресу www1.unece.org/stat/platform/display/QI/Quality+Indicators+Home.

C. Интеграция новых источников данных (таких, как большие данные) и традиционных источников

33. В настоящем разделе основное внимание уделяется интеграции новых источников данных (таких, как большие данные компаний мобильной связи, социальные сети, смарт-датчики, спутниковые изображения, веб-страницы, операции по кредитным картам и т.д.) с традиционными источниками официальной статистики.

1. Краткое описание экспериментов

34. Уже были проведены следующие эксперименты:

- тематические исследования по стратегиям веб-сбора данных: подходы к получению собранных с веб-сайтов данных для составления официальной статистики – тематическое исследование Новой Зеландии и обзора опыта других стран;
- измерение цен путем интеграции собранных с веб-сайтов данных: расчет и сопоставление индексов цен на основе различных источников больших данных в разных странах (Новая Зеландия);
- интеграция полученных с веб-сайтов данных для составления статистики цен (Венгрия); и
- интеграция потенциальных источников информации для подготовки статистических данных о вакантных рабочих местах (Венгрия).

35. Эти эксперименты, как правило, предусматривали следующие шаги:

- определение статистических рамок и границ работы;
- разработка и тестирование интеграции собранных с веб-сайтов данных и традиционных наборов статистических данных, включая извлечение и распознавание объектов и сопоставление объектов. Изучение новых методов увязки и сопоставления объектов (например, когда объект может иметь более расплывчатую структуру, чем запись); и
- оценка статистических продуктов, полученных в результате интеграции/комбинирования/объединения собранных с веб-сайтов данных и традиционных наборов статистических данных.

36. В 2016 году участники этого направления деятельности изучали текущую работу в данной области (в своих собственных организациях и в других группах (например, в рамках проектов по использованию больших данных ESSNET и ГВУ-МОС)). Эта деятельность будет продолжена в 2017 году с предложением:

- разработать практическое руководство по интеграции результатов обследований, административных данных и больших данных (включая тематические исследования) на основе работы, осуществляемой участвующими организациями;
- поощрять задействование других участников и проектов; и
- описать шаги, необходимые для использования ТМПСИ.

2. Соответствующая работа в рамках других проектов/организаций

37. Соответствующая работа, проводимая в международных организациях или в рамках других экспериментов проекта по интеграции данных ГВУ-МОС 2016 года, включает в себя:

- большие данные ЕЭК ООН, большие данные Евростата;
- работу в области цен (см. раздел B);
- расчет индекса арендной платы по Новой Зеландии на основе данных аукционного сайта Trade Me (без корректировки на качество); и
- вакантные рабочие места (Сербия).

D. Интеграция геопространственной и статистической информации

38. В настоящем разделе рассматриваются вопросы интеграции географической информации со статистической информацией.

39. Статистические данные практически всегда связаны с определенным физическим пространством, таким как муниципалитет, область, страна или регион. Каждый уровень является полезным для различных субъектов и различного рода решений. Многие из этих решений зависят от физических элементов окружающей среды и к тому же будут оказывать на них воздействие. Местоположение и объем природных ресурсов, типы почв, погодные условия, коммуникационная инфраструктура, производственные объекты служат примерами географической информации, которые являются необходимыми элементами для полного понимания цифр, предоставляемых официальной статистикой. Работа в области геопространственной классификации может стать отправной точкой для увязки статистики и географии.

40. Интеграция географических данных в официальную статистику призвана повысить полезность производимой статистической информации.

1. Краткое описание экспериментов

41. Ввиду сложности данного направления работы не предполагалось провести все мероприятия за один год. Проделанное послужило лишь началом работы, поскольку вполне вероятно, что будущая деятельность в этой области потребует усилий, которые выходят за рамки одной целевой группы. Эксперименты, связанные с этим типом интеграции данных, были сосредоточены на:

- составлении перечня уровней интеграции пространственных объектов, используемых в статистике и геодезии – «Модель уровня 10» для согласования систем пространственной привязки статистики и геодезии (Польша);
- анализе существующих моделей интеграции геопространственной и статистической информации (Колумбия); и
- интеграции информации об образовании на основе географического местоположения школ (Колумбия).

42. В рамках экспериментов были предприняты следующие шаги:

- проведен обзор международных усилий в этой области (например, проектов GEOSTAT и GISCO Евростата и ООН-УГГИ) с целью обеспечения синергизма и укрепления рабочего потенциала в рамках проекта;
- составлен перечень источников и проектов, связанных с географической информацией, которые могут иметь отношение к подготовке статистических данных;
- были разработаны предложения в отношении полезной информации, которая может производиться путем интеграции статистической и географической информации;
- были изучены методы интеграции и подготовки соответствующих информационных продуктов на основе комбинирования обоих видов данных; и
- были проведены эксперименты и пилотные работы по оценке полезности интегрированных статистических и географических данных с целью внесения корректив в сферу охвата этого проекта.

43. Были предложены следующие направления последующей деятельности на 2017 год:

- построение дерева решений – последовательности практических вопросов, на которые необходимо дать ответ до начала интеграции геопространственных данных со статистическими данными. Это поможет организациям оценить свою зрелость и потенциал в области пространственной статистики;

- изучение и рассмотрение вопроса о включении результатов работы участников проекта GEOSTAT 2 и Австралии по инвентаризации геопространственного потенциала и стандартов данных (включая стандарты геопространственного сообщества), в компоненты ТМПСИ и ТМСИ;
- изучение и рассмотрение возможностей совместной работы с ООН-УГГИ, Европой и Северной и Южной Америкой в целях достижения более эффективных результатов в области интеграции статистической и геопространственной информации;
- сопоставление и анализ результатов обследований, проведенных в различных регионах. Проведение дополнительных обследований на основе GEOSTAT 2, по мере необходимости, если методологии вышеупомянутых обследований являются отличными, а их результаты несопоставимыми. Анализ наличия общих черт и расхождений в подходах на национальном и международном уровнях к пространственным объектам, используемым в статистике и геодезии, в целях согласования двух систем пространственной привязки (реализация «Модели уровня 10» в нижних слоях GSGF); и
- выявление общих рисков для интеграции геопространственной и статистической информации.

2. Соответствующая работа в рамках других проектов/организаций

44. Бюро КЕС провело углубленный анализ темы «Развитие услуг в области геопространственной информации» в феврале 2016 года. Семинар КЕС на эту тему был проведен в апреле 2016 года, и Бюро обсудило последующие меры по итогам этих мероприятий в октябре 2016 года. На совещании Бюро в феврале 2017 года в рамках пункта III k) повестки дня было рассмотрено предложение о проведении совместной деятельности статистиков и геопространственного сообщества под эгидой Конференции.

45. Картина деятельности в области интеграции геопространственных и статистических данных выглядит весьма сложной и характеризуется участием многочисленных игроков во всем мире. Глобальная статистическая геопространственная модель (ГСГМ-ООН-УГГИ), Статистическая пространственная модель (СПМ), обеспечивающая интеграцию статистики и геопространственной информации в государствах – членах ЕС (ESSnet), Европейский форум по геодезии и статистике ЕФГС) и такие инициативы, как GEOSTAT 2 (Евростат), имеют чрезвычайно важное значение для разработки последовательного и систематического подхода к увязке геопространственных и статистических данных. Это, вероятно, потребует некоторого времени.

46. Ввиду сложности этой области было бы полезно провести очное или виртуальное экспресс-совещание ключевых игроков (включая Австралию, Польшу, Колумбию, Мексику, Финляндию, Швецию и других, которые будут определены позднее). Это экспресс-совещание планируется провести в ходе предлагаемого рабочего совещания, посвященного геопространственным и статистическим стандартам.

E. Проверка достоверности официальной статистики

47. Еще одной конкретной областью, в которой важную роль играет использование данных из других источников, является проверка достоверности официальной статистики. Поскольку интеграция данных позволяет выявлять записи в многочисленных источниках, относящиеся к одному и тому же лицу или единице, она может использоваться для проверки достоверности официальной статистики:

- либо путем использования внешних источников данных для определения точности результатов обследования;

- либо путем использования результатов обследований для проверки достоверности результатов из альтернативных источников поставщиков статистических данных.

48. Существуют примеры, когда другие источники считались сопоставимыми с официальной статистикой, и когда между ними возникали расхождения, под сомнение ставилась официальная статистика. Один пример (Соединенное Королевство) касается сопоставления совокупности предприятий, перечисленных в телефонных справочниках «Желтые страницы», с охватом статистического коммерческого регистра (www.unece.org/fileadmin/DAM/stats/documents/ces/sem.53/wp.7.e.pdf). Еще один пример касается сопоставления показателей инфляции, рассчитанных в рамках проекта «Миллиард цен» МТИ, с официальными индексами цен. Эти примеры свидетельствуют о том, что «другие источники» достигли уровня доверия, что способны бросить вызов официальной статистике. Специалисты по конъюнктурному анализу сталкиваются с аналогичными проблемами. Их бизнес уже подвергается усиливающемуся давлению со стороны дешевых или бесплатных интернет-групп.

49. Ясно одно. Эти «другие источники» возникают, чтобы остаться, и их число и влияние продолжат расти. По ряду ЦУР будут использоваться показатели, рассчитываемые на основе официальных источников, в которых могут не применяться стандартные методологии и источники данных. Такие различия в подходах могут привести к возникновению расхождений в результатах, которые потребуется проанализировать и разъяснить.

50. В рамках данного направления работы ведется изучение вопросов, связанных с интеграцией альтернативных источников данных в процессы проверки достоверности, используемые для подготовки официальной статистики. Эти вопросы включают в себя оценку происхождения и качества источника, надежность и коммерческие и иные интересы сторон, использующих их; процессы разработки и методы моделирования, которые являются устойчивыми и формализованными (поскольку специальные корректировки статистики будет трудно обосновать); обучение пользователей надлежащему использованию и толкованию информации (как широкой общественности, так и более конкретных групп пользователей).

51. В 2017 году планируется расширить руководящие принципы, разработанные в 2016 году, путем включения в них качественных показателей качества процесса проверки достоверности. Выводы должны быть сообщены участникам проекта ГВУ-МОС по разработке показателей качества для ТМПСИ.

1. Краткое описание экспериментов

52. Что касается использования интеграции данных для проверки достоверности официальной статистики, то в этой области было проведено два эксперимента:

- сравнительный анализ данных о доходах, полученных в рамках обследования доходов Новой Зеландии, и административных данных (Новая Зеландия); и
- увязка статистического регистра занятости и обследования рабочей силы (Словения).

53. Они предусматривали следующие шаги:

- выявление проблем, связанных с систематическим использованием данных из других источников в процессе проверки достоверности официальной статистики; и
- рекомендация потенциальных подходов и методов моделирования.

54. Работа в 2016 году была сосредоточена на выявлении различных приложений и методов для проверки достоверности официальной статистики. Выявленные проблемы, связанные с использованием административных данных для

проверки достоверности официальной статистики, и извлеченные из экспериментов уроки, а также другие проекты, осуществляемые в организациях участников, документируются с целью разработки первоначальных руководящих принципов использования административных данных для проверки достоверности официальной статистики, а также рекомендации подходов и методов моделирования для решения выявленных проблем.

55. Было предложено продолжить в 2017 году работу по:

- изучению релевантности использования Хранилища правил ESSNET и Справочника определений в области проверки достоверности ESSVIP в качестве инструментов проверки достоверности;
- тестированию рекомендуемых подходов и методов моделирования;
- описанию различных приложений, использующих интеграцию для проверки достоверности статистики (например, проверки достоверности имеющихся статистических данных, замены источников, разработки новых статистических данных, улучшения структуры имеющихся статистических данных, проверки достоверности результатов из альтернативных источников и поставщиков статистических данных);
- разработке первоначальных руководящих принципов использования административных данных для проверки достоверности официальной статистики, а также включению описания первоначального набора минимальных требований (в их число могут входить минимальные метаданные, этапы процесса, методы) для инициирования процесса проверки достоверности;
- определению того, какие методы проверки достоверности существуют или могут быть созданы, а также включению программного обеспечения, имеющегося для применения методов проверки достоверности;
- документированию проблем, выявленных при использовании административных данных для проверки достоверности официальной статистики;
- рекомендации подходов и методов моделирования, необходимых для решения проблем, выявленных при использовании административных данных для проверки достоверности официальной статистики;
- тестированию рекомендуемых подходов и методов моделирования;
- документированию опыта и извлеченных уроков;
- расширению первоначальных руководящих принципов использования административных данных для проверки достоверности официальной статистики, по мере необходимости, а также включению показателей качества или индикаторов, которые имеют важное значение для сообщения качества процесса проверки достоверности, и сообщению выводов участникам проекта ГВУ-МОС по разработке показателей качества для ТМПСИ; и
- разработке учебных материалов по реализации процесса проверки достоверности с использованием интеграции данных.

2. Соответствующая работа в рамках других проектов/организаций

56. В рамках ESSnet и ESS.VIP осуществляются следующие смежные проекты:

- Проект ESS.VIP ADMIN: этот проект направлен на поиск путей оптимизации использования и доступности источников административных данных в целях подготовки официальной статистики при одновременном гарантировании качества и сопоставимости этих статистических данных. Подробная информация о работе по этому проекту размещена по адресу

https://ec.europa.eu/eurostat/cros/content/essvip-admin-administrative-data-sources en;

- Проект ESSnet по интеграции данных: в рамках этого уже завершенного проекта основное внимание уделялось методологиям и методологическим вопросам интеграции микроданных. Подробная информация о работе по этому проекту размещена по адресу http://ec.europa.eu/eurostat/cros/content/data-integration-finished_en;
- Проект ESSnet по интеграции результатов обследований и административных данных: данный проект был направлен на расширение знаний и опыта участвующих НСУ в области использования интегрированных результатов обследований и административных данных для подготовки официальной статистики. Подробная информация о работе по этому проекту размещена по адресу http://ec.europa.eu/eurostat/cros/content/isad-finished_en;
- Проект ESSnet по макроинтеграции: в рамках этого проекта обсуждаются различные методы интеграции источников агрегированных данных или данных макроуровня. Подробная информация о работе по этому проекту размещена по адресу https://ec.europa.eu/eurostat/cros/content/macro-integration_en.

V. Возможности, проблемы и вызовы

A. Возможности и потенциальные выгоды

57. Интеграция данных из многочисленных источников данных позволяет НСУ расширять использование ими внешних источников данных в целях подготовки официальной статистики и обеспечивает многочисленные возможности и потенциальные выгоды. Эти факторы служат для статистических управлений мощными стимулами для укрепления потенциала в этой области. Интеграция данных способна:

- обеспечить разработку более своевременных и более подробных статистических данных;
- позволить создание новых видов официальной статистики или усовершенствовать существующие компоненты официальной статистики благодаря внешнему источнику данных;
- обеспечить удовлетворение новых и неудовлетворенных потребностей в данных;
- снизить нагрузку на респондентов;
- преодолеть последствия снижения коэффициентов предоставления ответов или заменить существующие источники данных; и
- повысить качество и решить проблемы погрешности в результатах обследований.

58. Интеграция данных позволяет также повысить качество традиционных статистических источников благодаря нахождению источников ошибок и определению методологических и практических проблем, которые влияют на качество. Новые источники данных могут обеспечить более эффективную выборку или даже полный охват. Они могут обходиться дешевле, чем результаты обследований, и потенциально обладать более высоким качеством. Онлайновые данные обладают преимуществами оперативности и высокой регулярности.

59. В Новой Зеландии, например, развитие экспертного потенциала в области интеграции данных привело к созданию комплексной инфраструктуры интегрированных данных (ИИД) Статистического управления, объединяющей увя-

занные наборы данных различных правительственных учреждений (включая собственные наборы данных Статистического управления). ИИД является крупной исследовательской базой данных, содержащей микроданные о лицах и домашних хозяйствах, объем которых постоянно растет. Она создала возможности для поиска ответов на сложные исследовательские вопросы, позволяющих повысить отдачу для новозеландцев.

60. Особым типом интеграции данных является интеграция административных данных и результатов обследований. Она может служить различным целям: дополнение результатов выборочных обследований по определенной части совокупности, по набору переменных, в целях оценки или для процесса проверки достоверности и редактирования данных. В некоторых случаях выборочное обследование может быть заменено данными, опирающимися исключительно на административные источники. Административные данные могут также служить источником для создания и ведения статистических регистров, которые затем будут использоваться для проведения обследований.

61. Выборочные обследования в целом носят более гибкий характер по сравнению с административными источниками, поскольку они предназначены для решения конкретной задачи. Административные источники, с другой стороны, обеспечивают, как правило, более широкий охват целевых совокупностей и, как правило, имеют высокие показатели предоставления ответов. Получение данных из существующих административных источников является более эффективным с точки зрения затрат и сопряжено с меньшими расходами по сравнению с проведением выборочного обследования и не создает никакой дополнительной нагрузки на респондентов. Поскольку административные данные охватывают всю совокупность, данные по малым районам могут быть рассчитаны на уровне детализации, который не позволяют выборочные обследования. Это также имеет преимущество с точки зрения осуществления политики на местном уровне. Однако существуют также и проблемы, связанные с административными источниками, о которых говорится в нижеследующем разделе.

62. Интеграция статистических данных и данных из других источников с геопространственной информацией повышает полезность как статистических, так и пространственных данных. Другие преимущества интеграции с геопространственными данными заключаются в:

- повышении качества геопространственных и статистических данных;
- укреплении сотрудничества между картографическими агентствами и статистическими управлениями, улучшении обновления и своевременности наборов данных;
- повышении функциональной совместимости наборов данных, упрощении методов увязки наборов данных;
- дополнительных возможностях (пространственного) анализа и представления данных;
- новых видах услуг и данных для удовлетворения потребностей пользователей;
- гибкости для внешних пользователей; и
- расширении использования: разработчиками политики и директивными органами, особенно региональными директивными органами; в научных целях; в целях охраны окружающей среды и т.д.

B. Вопросы и проблемы

63. Интеграция данных сопряжена с многочисленными проблемами. Термин «интеграция данных» может иметь широкое толкование: существует множество

различных типов данных и различных источников, которые могут подвергаться интеграции. Поэтому необходимы различные подходы.

1. Правовые и институциональные вопросы

64. Первым и наиболее важным правовым вопросом является правовая основа для использования административных и других внешних источников в статистических целях. Проекты интеграции данных, осуществляемые НСУ, подпадают под действие законодательства, кодексов практики, протоколов и политики, причем некоторые из которых изложены в Законе о статистике. Использование уже существующих административных источников в целях официальной статистики может быть предусмотрено национальным законодательством. В отсутствие такой основы необходимо ее создать.

65. Для использования других источников административных данных зачастую не требуется правой основы. Пятый основополагающий принцип официальной статистики гласит, что «Данные для статистических целей могут собираться из всех типов источников, будь то статистические обследования или административная отчетность. Статистические учреждения должны выбирать источник с учетом качества, своевременности, затрат и нагрузки на респондентов». Это дает НСУ право использовать данные из различных источников, но не обязывает владельцев других источников обмениваться своими данными со статистическими управлениями.

66. Законодательство по вопросам конфиденциальности и мнение общественности требуют особого учета при интеграции высоко детализированных административных данных, больших данных или пространственных данных.

67. Важно учитывать мнение общественности и вести тщательную информационную работу по вопросам интеграции данных в целях сохранения доверия к официальной статистике. Поощрение понимания и поддержки со стороны пользователей данных и широкой общественности в отношении работы по интеграции данных, проводимой в рамках официальной статистики, является важным вопросом. Для заручения поддержкой общественности необходимо соблюдать законы, касающиеся защиты данных, например, Закон о защите персональных данных. Они определяют правила обработки персональных данных таким образом, чтобы не допустить нарушения юридических прав граждан на неприкосновенность частной жизни и защиту персональных данных.

68. Необходимо наладить сотрудничество с поставщиками административных данных для обеспечения того, чтобы охват, качество данных, классификации и т.д., используемые в административных источниках, соответствовали статистическими требованиям. Сотрудничество НСУ с административными органами в деле подготовки правовых документов о создании и ведении административных источников является правильным подходом к решению этой задачи. Получение одобрения НСУ при принятии законодательства об административных данных может быть предусмотрено в законе о статистике. Возможно потребуется подписать соглашения о сотрудничестве для распределения обязанностей между сторонами соглашения, определения правил и условий передачи данных, такие как своевременность, техническая реализация и метаданные.

69. Интеграция данных открывает возможности для налаживания партнерств в целях взаимодействия и обмена опытом со статистическими учреждениями, разработчиками данных, академическими кругами, исследовательскими учреждениями, частным сектором, государственными учреждениями, коммерческими организациями, в том числе провайдерами инфраструктуры ИКТ. Эти партнерства сопряжены со всеми проблемами, возникающими при налаживании стратегических партнерств (которые были рассмотрены на семинаре КЕС в апреле 2016 года). В некоторых случаях одни и те же данные используются несколькими учреждениями, в связи с чем рекомендуется наладить постоянное сотрудничество в институциональных методологических группах для создания системы, удовлетворяющей и административным, и статистическим целям.

70. Переход от исследовательских проектов интеграции данных к тиражируемым и надежным процессам разработки статистики сопряжен со своими проблемами, включая управление проектами по интеграции данных. Ряд стран создали эффективные рамки для управления реализацией проектов интеграции данных. Например, Австралия имеет рамки для интеграции статистических данных («National Statistical Service § Data Integration Need to know,» 2014), которые были одобрены и используются в рамках австралийского правительства. В других странах имеются аналогичные рамки.

2. Управленческие вопросы, включая ресурсы

71. Общей проблемой в области интеграции данных является выделение надлежащих ресурсов: бюджета, людских ресурсов и ИТ-инфраструктуры для обеспечения необходимого уровня экспертизы, знаний и технологии. Бюджетные ограничения могут отрицательно сказываться на возможности получения необходимых ресурсов.

72. Под людскими ресурсами понимаются статистики, методологи и эксперты в области ИТ. Требуется определенное сочетание широкого круга различных компетенций как для работы с данными, так и для разрешения и согласования использования внешних источников данных и информирования о нем. К их числу могут относиться навыки лидерства, умение вести переговоры, налаживать отношения, компетенции в области защиты данных и коммуникации и т.д. Сотрудники, занимающиеся интеграцией данных должны быть осведомлены о различных стратегиях и законодательных нормах, касающихся интеграции данных, включая законодательство о защите персональных данных.

73. Основные компетенции включают в себя знания в области контента данных, статистического процесса, понимание показателей качества и т.д. Ключевое значение имеет наличие кадров с надлежащими компетенциями в области ИТ, программного обеспечения и баз данных. Конкретные проекты интеграции данных, такие как использование больших данных, могут требовать знания инструментов веб-сбора данных и информационной технологии для управления и обработки больших данных. Для использовании источников геопространственных данных необходим опыт работы с геопространственной информацией и знание соответствующих существующих процессов.

74. Интеграции данных способствует стремительный прогресс в области ИТ, т.е. аппаратных средств, а также широкое разнообразие программных средств. ИТ-инфраструктура, необходимая для интеграции данных, включает в себя серверы, программное обеспечение и инструменты создания баз данных для хранения микроданных и метаданных и инструменты обработки и распространения данных. Требуются безопасные и эффективные механизмы файлообмена. Для использования больших данных требуются специальные ИТ-ресурсы для хранения и обработки больших массивов данных. Геопространственные данные требуют дополнительных инфраструктурных ресурсов для того, чтобы данные сохранялись в дезагрегированном виде после первоначальной обработки до распространения. Эти услуги могут быть переданы на внешний подряд, и в этом случае необходимо обеспечить выделение необходимых средств на аутсорсинг.

75. Интеграция данных сопряжена с рисками, которыми необходимо управлять. Официальная статистика должна обеспечивать стабильный выпуск, в то время как исходные ресурсы для интеграции данных часто являются нестабильными. Поскольку существует вероятность перебоев в работе источников данных, необходимо предусмотреть план действий в чрезвычайных обстоятельствах, когда источник данных становится недоступным.

76. Для смягчения рисков следует использовать организационные рамки управления рисками и Руководящие принципы практического управления рисками в статистических организациях (в настоящее время разрабатываемые

ИСТАТ). Некоторые источники данных могут требовать особых подходов к оценке рисков.

77. Интеграция данных может быть сопряжена с высокими начальными затратами. Сопровождение действующего проекта интеграции данных также требует расходов. Поэтому можно рекомендовать сосредоточить сначала внимание на использовании ограниченных ресурсов для получения ощутимых и полезных результатов.

78. Доступ к внешним источникам данных не обязательно будет бесплатным. Кроме того, затрат требует и оценка качества внешних источников данных. Административные данные, как правило, обходятся дешевле результатов выборочных обследований, поскольку они уже собираются в административных целях, но они будут все же требовать определенного бюджета. Аналогичным образом использование больших данных требует соответствующих трудозатрат и бюджета для подготовки и экспериментирования с данными.

79. Различные потенциальные источники информации для интеграции данных могут требовать использования различных методов. Поэтому до запуска проекта по интеграции следует протестировать/изучить каждый источник данных. Все соответствующие расходы следует определить и оценить до начала реализации проекта по интеграции данных.

80. Еще одной проблемой является сопротивление изменениям любого компонента текущего процесса производства, которых потребует интеграция внешнего источника данных, особенно в тех случаях, когда существующие подходы пользуются широкой поддержкой и опираются на прочный кадровый фундамент.

81. Использование стандартных процессов, являющихся общими для различных типов интеграции данных, в значительной степени облегчит интеграцию данных. Многие проекты по интеграции данных могли бы использовать схожую последовательность шагов, такую как:

- выявление потребностей и уточнение бизнес-требований;
- изучение соответствующей работы, проводимой в других организациях;
- определение партнеров и сотрудничающих субъектов;
- отбор потенциальных источников данных;
- определение методологических соображений и соображений качества;
- анализ того, можно ли и каким образом будет использовать потенциальные источники данных (в качестве прямых источников или для проверки достоверности данных);
- подготовка экономического обоснования (включая издержки, выгоды, риски и т.д.);
- получение данных;
- получение необходимых инструментов, кадров, ресурсов;
- экспериментирование;
- оценка результатов;
- доработка методов и подходов (по мере необходимости); и
- разработка тиражируемого производственного решения.

82. Существует также ряд задач, для решения которых было бы полезно иметь стандартные процессы, такие как:

- редактирование расхождений в прошедших увязку данных;
- условный расчет значений переменной в прошедших увязку записях интегрированного набора данных;

- определение весов для корректировки на отсутствующие связи в рамках интегрированного набора данных; и
- применение методов проверки достоверности.

3. Методология, концепции и определения

83. Методология интеграции данных в значительной степени зависит от используемых источников данных, в связи с чем не существует каких-либо единых рекомендуемых методов для интеграции всех типов данных. Планируется создать каталог широко используемых и новых методов для различных типов интеграции данных в рамках проекта по интеграции данных ГВУ-МОС. В идеале методы должны описываться в такой форме, которая была бы совместима с Единой архитектурой статистического производства.

84. Большинство источников данных, используемых для интеграции, являются внешними по отношению к НСУ. НСУ не могут влиять на сбор данных для них, в связи с чем возможны различия в концепциях, классификациях, совокупностях и единицах сбора.

85. При интеграции данных статистические концепции следует согласовать с концепциями новых источников данных. В некоторых случаях придется разработать новые концепции. НСУ должно понимать подоплеку решений по конструкции данных, с тем чтобы оно могло определить, как преобразовать внешние данные в статистическую информацию. Эти различия сказываются на возможности использования внешних источников данных для подготовки статистики, конкретно с точки зрения: охвата совокупности, приложимости целевых концепций, доступности и точности дескриптивных метаданных, погрешности выборки, систематической ошибки, правовой основы разработки данных, методологии сбора данных/вопросника, нагрузки на респондентов, конфиденциальности результирующих продуктов и различных последствий для разных типов представленных данных. Эти различия должны быть четко разъяснены и задокументированы, и информация о них должна храниться для обеспечения повторного использования и совершенствования оценок.

86. Пользователи статистической информации должны быть хорошо информированы об определении концепций и совокупностей, используемых во всех источниках данных, применяемых в целях статистического производства. Также следует обеспечить единое понимание статистических концепций НСУ пользователями статистической информации.

87. Например, при интеграции административных данных и результатов обследований различия в концепциях могут привести к возникновению проблем охвата, а также проблем систематической ошибки. В некоторых случаях, таких как статистика предприятий, единицы, используемые в административных данных, не всегда соответствуют определению требуемых статистических единиц. Преобразование административных единиц в статистические потребует определенного моделирования. Вполне вероятно, что различия будут также иметься и в определениях переменных. Важно тщательно изучить влияние этих различий. В некоторых случаях можно будет повлиять на административное определение, сотрудничая с компетентным органом.

88. В случае различий в классификациях, обычным решением является использование таблиц соответствий и инструментов перекодировки на основе дополнительных переменных, которые могут иметься в наличии для перевода в более правильный классификационный код. Однако даже использование идентичных классификаций может приводить к получению разных данных, особенно в тех случаях, когда классификации носят сложный характер или правила классификации трудны для применения. В административных источниках зачастую производится кодирование респондента, в то время как вопросник выборочного обследования может содержать открытые вопросы, а кодирование во многих случаях осуществляться экспертами.

89. Сотрудничество между НСУ и административным органом является хорошим способом решения проблемы классификации. НСУ может предоставить свой опыт и отвечать за ведение классификации. Еще один вопрос заключается в том, следует или нет непосредственно использовать переведенные международные классификации или национальные классификации. Это зависит от того, для чего необходимы национальные данные. Первый вариант, как правило, сопряжен с большими трудностями в случае изменений и пересмотров по сравнению с национальными классификациями. Изменение классификации в административном источнике является сложной задачей, поскольку может существовать много поставщиков данных, которых необходимо будет ознакомить с изменениями.

90. При интеграции геопространственных данных с другими источниками уровень детализации или агрегирования наборов данных может не совпадать и охват также может быть различным.

91. Требующими решения проблемами являются также отсутствие данных и погрешности. В рамках статистических обследований отсутствие данных обусловлено непредставлением ответа по единице или по переменной, но в административных источниках причины могут быть иными. Важно определить, носят ли погрешности и отсутствие данных систематический характер, и применять надлежащие правила проверки достоверности и редактирования.

92. Общей проблемой увязанных наборов данных являются расхождения в увязанных записях. По мере увеличения число увязываемых наборов данных возрастает потенциал повышения эффективности работы по выявлению и устранению несоответствий в записях, поскольку увеличивается число переменных. Однако это может также вести к увеличению объема необходимого редактирования.

93. В отношении интегрированных наборов данных должны быть определены источники потенциального смещения. Проблемы охвата и концептуальные проблемы могут затрагивать только некоторые группы совокупности, поэтому следует проявлять осторожность при распространении результатов на всю совокупность. Некоторые переменные могут отрицательно влиять на качество увязки и служить источником возможного смещения при проведении анализа результирующих наборов данных. Может потребоваться изучение коэффициентов увязки различных подгрупп совокупности.

94. Стандартизация идентификаторов (или другой информации, позволяющей увязку статистической информации) в различных источниках является одним из наиболее важных аспектов интеграции административных данных. В отсутствие такой системы будет гораздо труднее проводить увязку различных источников и придется применять методы увязки и сопоставления данных.

95. Для определения соответствующих моделей учета ошибок увязки необходимы методы более точной оценки ошибок увязки. Ошибки увязки усиливают потенциальные ошибки охвата результирующей целевой совокупности. Осторожность следует также проявлять при создании статистических единиц на основе интегрированных наборов данных в тех случаях, когда один набор является внешним, поскольку единица во внешнем наборе данных может определяться иначе.

96. Крайнюю осторожность следует проявлять при перспективном и ретроспективном анализе увязанных данных, особенно в случае продольных данных. Лицо может иметь связь в одном квартале, но не в другом по причинам качества данных (или может иметь связь с другой записью). Использование увязанных наборов данных даже в целях проверки достоверности может привести к разрыву в ряде динамики, который придется устранять.

97. Данные из внешних источников могут страдать погрешностями измерения, например ошибкой достоверности, и эти ошибки могут распространяться в тех случаях, когда эти данные будут интегрироваться с другими источниками

данных для подготовки статистических материалов. Следовательно, необходимо обеспечить надлежащее понимание целевых концепций, используемых в наборах данных из внешних источников, прежде чем приступить к их применению для подготовки официальной статистики.

4. Качество

98. Качество как источников данных, так и составленных статистических данных должно измеряться, регулироваться и сообщаться. Оценки требуют следующие параметры качества: точность, релевантность, непротиворечивость, доступность, сопоставимость и своевременность. Важно иметь в наличии подробные дескриптивные метаданные для оказания помощи в оценке качества источников данных. Интеграция данных повышает роль метаданных. Должны быть определены минимальные и идеальные метаданные.

99. Необходимо будет разработать рамки качества для интегрированных данных в целях понимания «неопределенности» или надежности оценок, сопоставимости показателей как во времени, так и с аналогичными показателями, рассчитываемыми на основе традиционных источников. Чрезвычайно важное значение имеют рамки качества, которые призваны определять оптимальную конструкцию источника(ов) данных, способную свести к минимуму совокупное воздействие потенциальных ошибок на статистическую продукцию. Проект рамок качества для больших данных, подготовленный в рамках проекта ГВУ-МОС 2014 года по использованию больших данных, и традиционные системы качества могут использоваться в качестве основы для разработки предварительного проекта рамок качества для статистических данных составляемых на основе интегрированных источников данных.

100. При интеграции источников данных проблемами являются своевременность и различия в отчетном периоде. Административные или пространственные данные могут отставать по времени или не совпадать со статистическим отчетным период. Эта проблема может быть решена путем анализа влияния и, в случае необходимости, корректировки с помощью моделей.

5. Международное сотрудничество

101. В целях взаимного обмена знаниями и недопущения дублирования усилий между странами и организациями необходимо использовать коллективный опыт сообщества официальной статистики. Обмен опытом между организациями официальной статистики будет содействовать разработке и применению единых подходов в целях более действенного и эффективного включения интеграции данных в процессы статистического производства.

102. Было бы полезно создать форум для налаживания международного сотрудничества в целях решения общих проблем. В соответствующих случаях необходимо наладить координацию с международными группами.

103. Существует потенциал объединения усилий некоторых поставщиков данных и группы НСУ для изучения взаимных выгод и, возможно, разработки соглашений о предоставлении данных. При использовании внешних ресурсов ИТ управлениям следует наладить хорошие партнерские связи с разработчиками и администраторами таких систем.

104. В качестве руководящих инструментов следует использовать стандарты, разработанных под эгидой ГВУ-МОС, такие как ТМПСИ, ТМРСО и ТМСИ, и, когда это возможно, необходимо разрабатывать или предлагать услуги, совместимые с ЕАСП. Нам также необходимо определить другие общие стандарты и рамки, используемые в более широком контексте, которые имеют важное значение для задач интеграции данных.

105. Отсутствие четко установленных стандартов в области интеграции различных источников данных открывает возможность для предложения новых стандартов. Необходимо обеспечить соответствие между статистическими и

другими стандартами. Поэтому следует изучить стандарты, используемые в отраслях источников (например, образования, туризма, банковского дела).

106. Было бы полезно разработать единые подходы к конкретным областям статистики. Для этой цели могли бы использоваться тестовая среда и наборы выборочных данных ГВУ-МОС. НСУ из разных стран могли бы объединить свои усилия в целях выработки единого подхода к получению данных от многонациональных компаний.

VI. Выводы и рекомендации

107. По сравнению с традиционными подходами интеграция данных позволяет производить статистические данные более своевременно и с более высокой степенью детализации и регулярностью. Масштабы деятельности по интеграции данных будут только возрастать. С появлением все большего числа источников данных и ростом мощностей ИТ и инфраструктуры данных будет усиливаться необходимость интеграции различных источников.

108. В некоторых национальных управлениях уже накоплен опыт использования определенных типов интеграции данных. Во многих случаях этот опыт носит более ограниченный характер, и управления не располагают экспертными знаниями для работы со всеми типами интеграции данных. В области интеграции данных имеются ограниченные, если вообще существуют, руководящие принципы, как и всеобъемлющий обзор опыта. Для международного статистического сообщества может быть полезным налаживание сотрудничества и обмена опытом в области интеграции.

109. Также настоятельно рекомендуется продолжить реализацию проекта по интеграции данных ГВУ-МОС и расширить обмен опытом и активизировать сотрудничество между национальными статистическими управлениями и другими учреждениями, занимающимися разработкой официальной статистики, а также поставщиками данных. Обмен информацией и сотрудничество позволят добиться экономии средств и наладить взаимодействие в целях достижения прогресса на национальном и международном уровнях.

110. Реализация различных экспериментов по интеграции данных привела к разработке большого числа рекомендаций. Ниже приводится их краткий перечень:

- заручение поддержкой руководства высокого уровня и информирование руководства статистического управления о проекте и его целях и результатах;
- сохранение четкости целей и разъяснение важности проекта, спецификация характеристик проблемы, которую предстоит решить;
- налаживание эффективного сотрудничества с партнерами/поставщиками данных: четкое определение потребностей в данных, учет общих целей учреждений (сбор данных только один раз, сокращение излишних расходов), заключение соглашений;
- сотрудничество с пользователями данных;
- проведение консультаций с другими экспертами, поиск дополнительной информации о возможных методах и решениях для практического применения; обмен сотрудниками для приобретения опыта и изучения передовой практики;
- учет международных рекомендаций и стандартов в области статистики;
- обмен результатами;

- начало работы с выборки данных, изучение данных до начала эксперимента, сохранение терпения и настойчивости, рассмотрение широкого спектра решений; и
- измерение качества данных.

111. Что касается последней рекомендации относительно рамок качества данных, то более подробные рекомендации уже были подготовлены в рамках настоящего проекта. Они размещены на вики ГВУ-МОС по адресу http://www1.unece.org/stat/platform/x/axSzBw.