Обзор международного опыта использования больших данных в официальной статистике

Содержание

1. Обзор опыта использования данных сотовых операторов в официальной статистике в странах Европейского Союза

1. Италия

Италия является одним из лидеров в области интеграции данных сотовых операторов (далее – СО) в официальную статистику.

Период реализации: С 2015 года по настоящее время (несколько последовательных проектов).

Название проекта: «Benessere Equo e Sostenibile dei Territori» (BES).

Участники:

Национальный статистический институт: ISTAT (Istituto Nazionale di Statistica).

Сотовые операторы: Совместная работа с TIM (Telecom Italia Mobile) как основным партнером. В некоторых пилотах участвовали несколько операторов.

Другие участники: Министерство инфраструктуры и транспорта, академические институты (например, MIT).

Область исследования: Мобильность населения, маятниковая миграция, туризм, наличное население.

Основные результаты:

Измерение маятниковой миграции: ISTAT успешно интегрировал данные СО для расчета ежедневных потоков населения между муниципалитетами, что позволило создать детализированную карту трудовой и учебной мобильности. Это дало более точную картину, чем традиционные обследования.

Оценка наличного населения: Регулярная оценка населения, фактически находящегося на территории муниципалитетов в ночное и дневное время, что критически важно для планирования общественных услуг и транспорта.

Статистика туризма: Анализ данных о роуминге для оценки потоков иностранных туристов в Италию и между регионами.

Основные проблемы:

Репрезентативность: Обеспечение репрезентативности при работе с данными только одного оператора (хотя TIM - крупнейший).

Методология: Разработка сложных алгоритмов для фильтрации шума, различения жителей и посетителей, учета нескольких SIM-карт у одного пользователя.

Конфиденциальность: Строгое соблюдение Акта о защите данных (далее – GDPR), агрегация и анонимизация данных на ранних этапах обработки.

Планы на продолжение: ISTAT планирует сделать статистику на основе CO регулярной и расширить ее применение для мониторинга целей устойчивого развития (SDGs) и анализа воздействия чрезвычайных ситуаций (например, пандемии COVID-19).

2. Нидерланды

Статистичекое управление Нидерландов (CBS) реализует один из самых зрелых проектов в ЕС.

Период реализации: С 2016 года по настоящее время.

Название проекта: «Statistics Netherlands Mobility Model».

Участники:

Национальный статистический институт: Statistics Netherlands (CBS).

Сотовые операторы: Партнерство с крупнейшими операторами страны (KPN, VodafoneZiggo). CBS имеет доступ к агрегированным данным от всех основных операторов.

Другие участники: Министерство инфраструктуры и водного хозяйства.

Область исследования: Мобильность, транспорт, статистика населения, туризм.

Основные результаты:

Официальная статистика наличного населения: CBS первым в ЕС начало публиковать ежеквартальные оценки наличного населения (день/ночь) на муниципальном уровне, используя CO. Эти данные используются для корректировки традиционных демографических прогнозов.

Детальная статистика мобильности: Создание модели мобильности, которая показывает потоки людей между муниципалитетами в разные дни и время, используется для планирования транспортной инфраструктуры.

Измерение туризма: Оценка количества туристов в конкретных районах (например, в районе Амстердама), их страны происхождения (на основе роуминга) и продолжительности пребывания.

Основные проблемы:

Юридические аспекты: Получение правового мандата на сбор данных от операторов. CBS действует на основе специального закона, обязывающего предоставлять данные для официальной статистики.

Качество данных: Тщательная валидация и калибровка данных CO с помощью традиционных источников (перепись, регистры).

Планы на продолжение: Интеграция CO с другими источниками данных (данные GPS транспорта, социальные медиа) для создания комплексной картины мобильности и динамики населения.

3. Португалия

Период реализации: Пилотный проект запущен в 2017-2018 гг., с продолжением и развитием.

Название проекта: «Big Data for Tourism Statistics» (в сотрудничестве с Евростатом).

Участники:

Национальный статистический институт: Statistics Portugal (INE).

Сотовые операторы: Партнерство с основными операторами страны (NOS, MEO, Vodafone Portugal). INE удалось обеспечить сотрудничество с тремя крупнейшими операторами.

Другие участники: Евростат (методологическая и финансовая поддержка), Banco de Portugal.

Область исследования: Туризм (въездной и внутренний).

Основные результаты:

Высокочастотная статистика туризма: Проект позволил оценить количество туристов в режиме, близком к реальному времени, с детализацией по регионам и странам происхождения.

Валидация традиционных методов: Данные СО использовались для проверки и дополнения данных опросов и регистров размещения (например, о количестве однодневных посетителей, которые не учитываются в отелях).

Анализ туристических потоков: Визуализация перемещений туристов внутри страны, выявление популярных маршрутов и точек интереса.

Основные проблемы:

Идентификация туристов: Разработка алгоритмов для точного различения туристов от местных жителей и трудовых мигрантов.

Многооператорность: Сложность агрегации и согласования данных от нескольких операторов для получения общей картины без дублирования.

Планы на продолжение: INE рассматривает возможность включения статистики на основе СО в регулярную производственную статистику по туризму.

4. Венгрия

Период реализации: С 2018 года (пилотные проекты).

Название проекта: «Mobile Position Data for Official Statistics».

Участники:

Национальный статистический институт: Hungarian Central Statistical Office (HCSO).

Сотовые операторы: Партнерство с одним из ведущих операторов (предположительно, Magyar Telekom).

Другие участники: Университеты и исследовательские центры (например, Центр экономических и региональных исследований).

Область исследования: Наличное население, маятниковая миграция, туризм.

Основные результаты:

Карта наличного населения: Создание детализированных карт плотности населения в дневное и ночное время для Будапешта и других крупных городов.

Анализ трудовой миграции: Количественная оценка потоков маятниковых мигрантов между округами.

Оценка туристской активности: Анализ данных для оценки нагрузки на туристические центры в Будапеште и на озере Балатон.

Основные проблемы:

Методологическая гармонизация: Приведение методологии в соответствие с требованиями Евростата.

Правовая база: Адаптация национального законодательства для легитимизации сбора и обработки СО для статистических целей.

Планы на продолжение: Развитие методологии для регулярного производства статистики туризма и мобильности и интеграции данных в систему демографической статистики.

5. Латвия

Период реализации: 2020-2021 гг. (пилотный проект).

Название проекта: «Use of Mobile Network Data for Official Statistics» (в рамках программы Евростата ESSnet Big Data).

Участники:

Национальный статистический институт: Central Statistical Bureau of Latvia (CSB).

Сотовые операторы: Работа с ведущими операторами (LMT, Tele2).

Другие участники: Евростат.

Область исследования: Туризм, мобильность.

Основные результаты:

Пилотная статистика туризма: Оценка потоков иностранных туристов в Латвию и их перемещений между регионами.

Оценка мобильности в Риге: Анализ ежедневных потоков людей внутри столичного региона.

Методологический задел: Разработаны внутренние протоколы и алгоритмы обработки CO.

Основные проблемы:

Недостаток опыта: Необходимость быстрого обучения сотрудников CSB.

Техническая инфраструктура: Создание платформы для обработки больших объемов данных.

Планы на продолжение: Оценка результатов пилота и принятие решения о переходе к регулярному производству.

6. Франция

Период реализации: Исследовательские и пилотные проекты с 2015-2016 гг., но без перехода в регулярную официальную статистику.

Название проекта: Проекты под эгидой Национального института статистики и экономических исследований (INSEE) и Управления генерального комиссара по устойчивому развитию.

Участники:

Национальный статистический институт: INSEE.

Сотовые операторы: Ограниченное сотрудничество с операторами (Orange) в рамках строгих исследовательских соглашений.

Другие участники: Агломерации (например, Greater Paris), исследовательские институты (L'Institut Paris Region).

Область исследования: Маятниковая миграция, мобильность в Ile-de-France, наличное население.

Основные результаты:

Исследовательские отчеты: Были опубликованы детальные анализы мобильности в парижском регионе, валидирующие и дополняющие данные традиционных опросов.

Методологические наработки: INSEE разработал прототипы методологий обработки СО, но не внедрил их в производство.

Основные проблемы:

Строгое регулирование: Франция имеет одно из самых строгих в ЕС законодательств о защите данных (CNIL). Получение разрешения на использование персональных данных, даже агрегированных, является крайне сложным.

Правовой мандат: У INSEE отсутствует прямой мандат на принудительный сбор данных у коммерческих компаний, в отличие от CBS в Нидерландах.

Планы на продолжение: Продолжение методологических исследований. Переход к производственной статистике маловероятен без изменений в законодательстве или появления готовых агрегированных продуктов от операторов.

7. Германия

Период реализации: Отдельные академические и прикладные исследования, но практически полное отсутствие проектов под эгидой Федерального статистического ведомства (Destatis).

Участники:

Национальный статистический институт: Destatis (не активен).

Сотовые операторы: Телекомы (Telefónica Germany/O2) активно сотрудничают с научными и коммерческими организациями.

Другие участники: Научно-исследовательские центры (Немецкий центр аэрокосмических исследований - DLR), коммерческие аналитические компании, федеральные и земельные министерства транспорта.

Область исследования: Транспортное планирование, анализ мобильности, туризм (на региональном уровне).

Основные результаты:

Данные используются для анализа транспортных потоков на автомагистралях, оценки заполняемости курортов, но не для официальной статистики населения или туризма.

Проект «MDM” (Mobility Data Marketplace): Платформа для обмена данными о мобильности, где СО являются одним из источников, но опять же не для Destatis.

Основные проблемы:

Правовой барьер: Интерпретация GDPR в Германии крайне строгая. Использование СО без явного и информированного согласия пользователя считается практически невозможным.

Общественное мнение: Высокая чувствительность населения к вопросам слежки и защиты данных.

Планы на продолжение: Destatis не аноНССровал планов по использованию СО. Развитие возможно только в рамках узкоспециализированных, строго регламентированных исследовательских проектов.

8. Испания

Период реализации: Пилотные проекты с 2018 года.

Название проекта: «Big Data para el Turismo» (Национальный институт статистики - INE), «Spanish Big Data for Official Statistics» (в рамках ESSnet Big Data).

Участники:

Национальный статистический институт: INE.

Сотовые операторы: Collaboration with major operators (Telefónica, Orange, Vodafone).

Другие участники: Министерство промышленности, торговли и туризма, региональные власти.

Область исследования: Туризм (въездной и внутренний), мобильность.

Основные результаты:

Высокочастотный мониторинг туризма: Оценка количества туристов в основных туристических зонах (Каталония, Балеарские и Канарские острова) с недельной или даже ежедневной периодичностью.

Анализ поведения туристов: Картирование маршрутов перемещений, продолжительности пребывания, выявление зон интереса.

Основные проблемы:

Согласование данных от нескольких операторов: Устранение дублирования и создание единой панели данных.

Сезонность: Разработка стабильных алгоритмов, работающих как в пик сезона, так и в межсезонье.

Планы на продолжение: INE работает над интеграцией данных CO в систему официальной статистики туризма в качестве дополнения к традиционным опросам.

9. Финляндия

Период реализации: С 2017 года (пилоты), активная фаза с 2020 года.

Название проекта: «Finergeo» (коммерческий продукт), сотрудничество с Statistics Finland.

Участники:

Национальный статистический институт: Statistics Finland.

Сотовые операторы: Партнерство с оператором DNA (данные агрегируются через платформу Finergeo).

Другие участники: Компания Finergeo (поставщик данных), муниципалитеты, Министерство транспорта и коммуникаций.

Область исследования: Наличное население, мобильность, градостроительство, туризм.

Основные результаты:

Статистика для муниципалитетов: Statistics Finland использует данные Finergeo для предоставления муниципалитетам информации о фактическом населении и потоках в режиме, близком к реальному времени.

Анализ эффективности общественного транспорта: Оценка пассажиропотоков и выявление узких мест.

Основные проблемы:

Коммерческая модель: Зависимость от стороннего поставщика агрегированных данных, а не прямое сотрудничество с операторами.

Репрезентативность: Охват только клиентов одного оператора.

Планы на продолжение: Расширение использования данных для мониторинга региональной доступности услуг и анализа воздействия политики в области транспорта.

10. Польша

Период реализации: С 2019 года.

Название проекта: «System for Providing, Collecting and Sharing Mobility Data» (национальный проект).

Участники:

Национальный статистический институт: Statistics Poland (GUS) — участвует как заинтересованная сторона.

Сотовые операторы: Планируется участие всех крупных операторов (Orange, T-Mobile, Play).

Координатор: Министерство развития и технологий.

Область исследования: Мобильность, транспорт, городское планирование, кризисное управление.

Основные результаты:

Создание национальной системы: Проект находится на стадии разработки инфраструктуры и нормативной базы. Цель — создать централизованную платформу для агрегации анонимных данных СО от всех операторов.

Пилоты для городов: Проводятся локальные пилоты в крупных городах (Варшава, Краков) для анализа транспортных потоков.

Основные проблемы:

Масштаб проекта: Создание общегосударственной системы — сложная организационная и техническая задача.

Координация между ведомствами: Необходимость согласования интересов Минразвития, GUS, операторов связи и органов местного самоуправления.

Планы на продолжение: Запуск полнофункциональной национальной системы сбора и анализа СО в 2026 году.

11. Эстония

Период реализации: Пилотные проекты с 2020 года.

Название проекта: Проекты под эгидой Департамента статистики (Statistikaamet).

Участники:

Национальный статистический институт: Statistics Estonia.

Сотовые операторы: Партнерство с крупнейшим оператором Elisa.

Другие участники: Министерство экономики и коммуникаций.

Область исследования: Туризм, мобильность.

Основные результаты:

Пилот по статистике туризма: Успешная оценка потоков туристов в Таллине и основных туристических зонах, сравнение с данными из регистров средств размещения.

Анализ трансграничной мобильности: Изучение ежедневных потоков между Эстонией и Латвией/Финляндией.

Основные проблемы:

Малое население: Риски деанонимизации при работе с детальными данными в малонаселенных районах.

Охват: Данные одного оператора могут быть недостаточны для создания полной картины.

Планы на продолжение: Интеграция данных СО в систему принятия решений на муниципальном и национальном уровне, особенно в сфере туризма.

12. Великобритания (не входит в ЕС, но значимый пример)

Период реализации: Активно с 2020 года в рамках реагирования на пандемию.

Название проекта: «COVID-19 Mobility Data Network», проекты Управления национальной статистики (ONS).

Участники:

Национальный статистический институт: Office for National Statistics (ONS).

Сотовые операторы: Партнерство с O2 (Telefónica UK), Vodafone, BT.

Другие участники: Департамент транспорта, научные консультанты (Университета Лидса и др.).

Область исследования: Мобильность во время пандемии, наличное население, туризм.

Основные результаты:

Мониторинг локдаунов: ONS и Департамент транспорта публиковали еженедельные отчеты об изменениях в мобильности населения, что было критически важно для оценки эффективности ограничений.

Оценка наличного населения в Лондоне: Анализ изменений в численности населения столицы во время и после пандемии.

Основные проблемы:

Срочность: Проекты запускались в авральном режиме, что не позволило провести полноценную методологическую проработку.

Интерпретация данных: Сложности в различении добровольного снижения мобильности от вызванного ограничениями.

Планы на продолжение: ONS намерен институализировать использование СО для постоянного мониторинга мобильности и популяционной статистики в крупных городских агломерациях.

Кросс-страновой анализ проблем и вызовов

1. Методологические проблемы:

Репрезентативность: Данные от одного оператора могут быть смещены.

Идентификация и классификация: Разработка алгоритмов для различения жителей, приезжих, трудовых мигрантов, туристов, определения места жительства и работы.

Агрегация и анонимизация: Необходимость работать с агрегированными данными, что ограничивает глубину анализа, но гарантирует конфиденциальность.

2. Правовые и этические проблемы:

Соответствие GDPR: Наиболее значительная преграда. Любая обработка СО должна быть анонимной, для четко определенных целей и с надлежащим правовым основанием.

Доверие общества: Необходимость прозрачности и коммуникации с общественностью о том, как используются их данные.

3. Технические и ресурсные проблемы:

Вычислительная инфраструктура: Обработка терабайтов данных требует мощных серверов и облачных технологий.

Квалификация кадров: Необходимость найма или переобучения статистиков в области науки о данных.

4. Институциональные проблемы:

Сотрудничество с операторами: Операторы не всегда заинтересованы в сотрудничестве из-за коммерческой ценности данных и рисков для репутации.

Гармонизация на уровне ЕС: Отсутствие единой методологии затрудняет сравнение данных между странами.

Общие выводы и тенденции

Использование данных сотовых операторов в официальной статистике ЕС перешло из стадии теоретических исследований в стадию активных пилотов и, в случае лидеров (Италия, Нидерланды), в стадию регулярного производства. Основные тренды:

1. От пилотов к производству: Успешные проекты доказывают свою ценность и интегрируются в регулярный статистический процесс.

2. Фокус на население и мобильность: Наиболее зрелые применения связаны с оценкой наличного населения, туризма и анализом перемещений.

3. Рост кооперации: Евростат играет ключевую роль в финаНССровании пилотов, обмене опытом и разработке гармонизированных методологий (например, в рамках проекта ESSnet Big Data).

4. Усиление правового поля: НСС и Евростат активно работают над созданием надежных правовых рамок, соответствующих GDPR.

5. Комплексный подход: СО все чаще используются не изолированно, а в сочетании с другими источниками (спутниковые данные, данные транспорта, социальные медиа) для получения более полной картины.

Модели взаимодействия НСС и сотовых операторов

Анализ по всем рассмотренным странам позволяет выделить несколько моделей взаимодействия между НСС и сотовыми операторами.

1. Модель на основе государственного мандата (наиболее эффективная)

Страна-пример: Нидерланды.

Механизм: Специальный закон («Закон о статистике») наделяет CBS правом обязать компании предоставлять данные, необходимые для производства официальной статистики. Это не покупка, а исполнение юридической обязанности.

Условия: Строгое соблюдение конфиденциальности: данные обезличены, агрегируются и используются исключительно для статистических целей. Операторы несут административную ответственность за непредоставление данных.

Преимущества: Стабильность, полнота данных (от всех операторов), отсутствие затрат для НСС.

Недостатки: Требует сильного правового основания и высокого уровня доверия к государственным институтам.

2. Модель добровольного партнерства (наиболее распространенная)

Страны-примеры: Италия, Португалия, Испания, Финляндия, Эстония, Венгрия, Латвия.

Механизм: НСС заключает соглашения о сотрудничестве с одним или несколькими операторами на добровольной основе. Часто это оформляется

как «пилотный проект» или «исследование» с четко оговоренными целями и условиями конфиденциальности.

Коммерческая основа:

Бесплатно (наиболее часто): Операторы предоставляют данные бесплатно, рассматривая это как корпоративную социальную ответственность, вклад в развитие государства и способ укрепления отношений с регулятором.

Оплата затрат: НСС может компенсировать оператору прямые затраты, связанные с подготовкой и передачей специально сформированных агрегированных наборов данных.

Коммерческий контракт (редко): В случае привлечения посредника (как Finergeo в Финляндии) НСС покупает уже готовые агрегированные данные или аналитические отчеты.

Преимущества: Гибкость, возможность быстрого старта пилотов.

Недостатки: Нестабильность (оператор может выйти из соглашения), риск отсутствия данных от всех операторов, потенциальная высокая коммерческая стоимость в будущем.

3. Модель национальной инфраструктуры (перспективная)

Страна-пример: Польша.

Механизм: Государство инициирует создание централизованной платформы, куда операторы по закону или соглашению передают агрегированные данные. Доступ к платформе получают уполномоченные госорганы, включая НСС.

Коммерческая основа: Финансируется из госбюджета (создание платформы). Операторы могут получать компенсацию затрат или предоставлять данные в обмен на доступ к агрегированной аналитике.

Преимущества: Всеобъемлющий охват, стандартизация, эффективность.

Недостатки: Высокая стоимость и сложность реализации.

4. Модель правового ограничения (тупиковая для официальной статистики)

Страны-примеры: Германия, Франция.

Механизм: Строгое законодательство о защите данных де-факто блокирует передачу данных от операторов НСС без практически невыполнимых условий (например, явного согласия каждого абонента).

Коммерческая основа: Отсутствует. Проекты заморожены на стадии теоретических обсуждений.

Итог: Официальная статистика в этих странах лишена этого инструмента, что создает для них конкурентное отставание по сравнению с лидерами.

Итоговый вывод: Успешность интеграции СО в официальную статистику напрямую зависит от способности государства создать сбалансированную правовую среду, которая, с одной стороны, защищает приватность граждан, а с другой - предоставляет НСС мандат или возможность для получения данных в интересах общества. На текущий момент в странах ЕС именно правовой фактор, а не технический или методологический, является ключевым определяющим элементом.

Нормативное правовое регулирование использования данных сотовых операторов в официальной статистике ЕС

В отличие от ДЗЗ, на уровне ЕС отсутствует прямое и единое законодательство, наделяющее статистические службы мандатом на сбор данных СО. Регулирование формируется в рамках общего правового поля защиты данных и конфиденциальности.

На уровне ЕС (Общеправовая рамка):

1. General Data Protection Regulation (GDPR) - Общий регламент по защите данных (Регламент (ЕС) 2016/679):

Ключевое значение: Это главный правовой барьер и одновременно основа для легитимизации проектов.

Статья 6: Требует правового основания для обработки персональных данных. Для статистики используется основание «выполнение задачи, осуществляемой в общественных интересах» (ст. 6(1)(e)), что должно быть закреплено в национальном законе.

Статья 89: Позволяет обрабатывать персональные данные для статистических целей при условии соблюдения принципов «минимизации данных» (data minimization) и «обезличивания» (pseudonymisation). Это

означает, что НСС могут получать только агрегированные и строго анонимизированные данные, а не данные о конкретном абоненте.

2. Регламент (ЕС) 2019/2152 о европейской статистике бизнеса:

Ключевое положение: Как и в случае с ДЗЗ, он призывает использовать альтернативные источники, но не дает прямого разрешения СО на доступ к коммерческим данным. Он создает общую обязанность по инновациям, но не преодолевает барьер GDPR в отсутствие национальных мандатов.

На национальном уровне (Ключевой уровень регулирования):

Поскольку телекоммуникационный сектор и детальное статистическое законодательство — это сфера национальной компетенции, именно на национальном уровне принимаются ключевые решения.

Модель 1: Явный государственный мандат (наиболее редкая).

Пример: Финляндия. Закон «О Статистике Финляндии» прямо наделяет Statistics Finland правом запрашивать данные у сотовых операторов для производства официальной статистики. Это позволяет работать в парадигме обязательного предоставления данных.

Модель 2: Отсутствие прямого мандата, работа в рамках GDPR (наиболее распространенная).

Пример: Нидерланды, Италия. НСС не имеют права требовать данные. Они вынуждены работать в модели добровольного партнерства. В этом случае правовым основанием является не государственный мандат, а «законный интерес» (legitimate interest, ст. 6(1)(f) GDPR) оператора или согласие, что гораздо менее устойчиво. Вся цепочка обработки должна быть выстроена так, чтобы гарантировать анонимность и агрегированность данных на самом раннем этапе.

Модель 3: Правовой вакуум (де-факто запрет).

Пример: Германия и Франция. Строгое толкование GDPR и законов о телекоммуникациях делает любые проекты по обмену данными СО с государственными органами, включая НСС, практически невозможными без изменения федерального законодательства.

Вывод: В отличие от ДЗЗ, где существует четкая общеевропейская правовая и финансовая поддержка, для СО правовая база фрагментирована и в основном носит запретительный характер. Прогресс достигается там, где

национальные законы создают для НСС исключения из общих правил защиты данных.

Обзор международного опыта использования данных ДЗЗ в официальной статистике

1. Место данных ДЗЗ в классификации больших данных ООН

В отличие от данных СО, данные ДЗЗ относятся к принципиально иной категории в классификации ООН:

Категория: «Данные, генерируемые машинами» (Machine-Generated Data)

Подкатегория: «Спутниковые снимки и дистанционное зондирование» (Satellite Imagery and Remote Sensing)

Ключевое отличие: Данные ДЗЗ являются объективными измерениями физических характеристик Земли, а не продуктом человеческой деятельности. Они пассивно фиксируют состояние окружающей среды.

2. Источники данных и способы доступа

Источники данных:

Спутники гражданского назначения:

Бесплатные (Open Data): Программа Copernicus (ЕС, спутники Sentinel-1, -2, -3 и др.), Landsat (NASA/USGS), MODIS (NASA). Являются основным источником для большинства статистических проектов.

Коммерческие: Planet Labs, Airbus (Pleiades), Maxar (WorldView). Обеспечивают более высокое разрешение, но за плату.

Беспилотные летательные аппараты (БПЛА): Для съемки с очень высоким разрешением на локальных территориях.

Авиационная съемка.

Способы доступа:

Прямое скачивание с порталов: Через API или порталы (например, Copernicus Open Access Hub, USGS EarthExplorer).

Облачные платформы: Доступ к данным и вычислительным мощностям напрямую в облаке, что становится стандартом де-факто. Ключевые платформы:

Google Earth Engine (GEE)

Microsoft Planetary Computer

CREODIAS (для данных Copernicus)

Сервисы Value-Added: Покупка уже готовых продуктов (например, карт классификации землепользования) у коммерческих провайдеров.

3. Сферы применения в официальной статистике

Данные ДЗЗ используются во многих отраслях статистики, связанных с окружающей средой, сельским хозяйством, землепользованием, мониторингом ЦУР.

1. Общеевропейские проекты и инфраструктура (координируемая Евростатом)

Период: С 2018 года по настоящее время (активная фаза).

Название проекта/Платформа:

ESS Copernicus Project (Европейская статистическая система и Copernicus)

Platform for Big Data in Official Statistics (EDP - Esper Big Data)

Цели: Создание общеевропейской инфраструктуры и методологии для широкого использования данных ДЗЗ в производстве официальной статистики, снижение нагрузки на респондентов, повышение детализации, точности и своевременности данных.

Участники:

Координатор: Евростат.

Национальные статслужбы: Все НСС стран-членов ЕС.

Космическое агентство: Европейское космическое агентство (ЕКА).

Сервисные компании: Поставщики данных и сервисов Copernicus (например, Sinergise, Sentinel Hub).

Область исследования: Сельское хозяйство, землепользование, экология, энергетика, транспорт.

Основные результаты:

Создание EDP (Esper Big Data Platform): Единой облачной платформы ЕС для обработки больших данных, включая спутниковые снимки.

Разработка пилотных модулей: Готовые алгоритмы и рабочие процессы (pipelines) для расчета конкретных показателей.

Интеграция с LUCAS (Land Use and Cover Area frame Survey): Объединение данных наземных обследований LUCAS со спутниковыми снимками для автоматической классификации землепользования и растительного покрова на всей территории ЕС.

Основные проблемы:

Вычислительные мощности: Обработка петабайтов данных с спутников Sentinel требует огромных ресурсов.

Стандартизация методологии: Обеспечение сопоставимости результатов, полученных разными НСС.

Необходимость валидации: Обязательная проверка результатов алгоритмов в сравнении с данными традиционных обследований (например, LUCAS, IACS).

Планы на развитие: Полная интеграция данных ДЗЗ в регулярное производство статистики к 2025-2027 годам. Расширение на новые домены: мониторинг городской среды, биоразнообразия.

2. Сельскохозяйственная статистика

Период: С 2020 года, активное внедрение.

Название проекта: Census of Agriculture 2020 с использованием ДЗЗ, Оценка урожайности и состояния посевов.

Цели: Картографирование сельскохозяйственных угодий, идентификация культур, оценка урожайности, мониторинг выполнения условий Общей сельскохозяйственной политики (CAP).

Участники: Евростат, НСС стран ЕС (например, ISTAT (Италия), Destatis (Германия), INSEE (Франция)), ЕКА.

Область исследования: Сельское хозяйство.

Связь с ЦУР: ЦУР 2.4 (устойчивые системы производства пищи), ЦУР 15.1 (сохранение экосистем).

Основные результаты:

Создание автоматизированных карт землепользования и севооборотов с разрешением 10x10 метров (данные Sentinel-2).

Ранняя оценка урожайности ключевых культур (пшеница, кукуруза, рапс) на основе вегетационных индексов (NDVI).

Мониторинг практик «озеленения» (Greening) в рамках CAP, например, наличие участков с разнотравьем.

Основные проблемы:

Сложность классификации культур со схожими спектральными сигнатурами (например, ячмень и пшеница).

Влияние облачности на оптические снимки (решается использованием радарных данных Sentinel-1).

Необходимость интеграции с административными данными (IACS - Integrated Administration and Control System) для верификации.

Планы на развитие: Переход к созданию «Цифрового двойника» сельского хозяйства ЕС для прогнозного моделирования и оценки воздействия климатических изменений.

3. Экологическая статистика и мониторинг ЦУР

Период: Постоянно, с резким увеличением роли ДЗЗ в последние 5 лет.

Название проекта: Мониторинг ЦУР 6 (вода), 11 (города), 13 (климат), 15 (экосистемы).

Цели: Оценка качества водных ресурсов, мониторинг урбанизации и "тепловых островов", отслеживание выбросов парниковых газов, картографирование лесов и деградации земель.

Участники: Евростат, Объединенный исследовательский центр (JRC), Европейское агентство по окружающей среде (EEA), НСС.

Область исследования: Экология, изменение климата, урбанистика.

Связь с ЦУР: Прямой мониторинг ЦУР 6, 11, 13, 15.

Основные результаты:

ЦУР 11.3.1: Расчет индекса соотношения темпов потребления земли к темпам роста населения на основе классификации искусственных поверхностей по снимкам.

ЦУР 15.3.1: Мониторинг доли деградированных земель на основе данных о продуктивности растительного покрова и землепользовании.

ЦУР 6.6.1: Отслеживание изменений в водных экосистемах (озера, реки, водохранилища).

Оценка концентрации NO2 и CO2 в атмосфере над городами и промышленными центрами (данные Sentinel-5P).

Основные проблемы:

Сопоставимость на глобальном уровне: Различия в методах расчета одних и тех же показателей ЦУР в разных странах.

Требуемая точность: Для некоторых экологических показателей разрешения спутников Sentinel может быть недостаточно.

Планы на развитие: Автоматизированное производство отчетов по ЦУР с высокой пространственно-временной детализацией. Использование ИИ для обнаружения изменений в режиме почти реального времени.

4. Энергетика и транспорт

Период: Пилотные и исследовательские проекты.

Название проекта: Оценка производства солнечной энергии, мониторинг транспортных потоков.

Цели: Картографирование установок возобновляемой энергетики, оценка интеНССвности судоходства и использования аэропортов.

Участники: Евростат, JRC, национальные министерства.

Область исследования: Энергетика, транспорт.

Связь с ЦУР: ЦУР 7.2 (возобновляемая энергия), ЦУР 9.1 (инфраструктура).

Основные результаты:

Обнаружение и классификация солнечных панелей на крышах зданий с помощью нейронных сетей (CNN) на основе спутниковых снимков высокого разрешения.

Основные проблемы:

Для энергетики: Сложность оценки реальной выработки энергии только по снимкам.

Для транспорта: Необходимость интеграции множества источников данных.

Планы на развитие: Создание регулярной статистики по распространению и потенциалу малой солнечной энергетики. Мониторинг загруженности логистических цепочек.

Ключевые выводы по ЕС:

1. Системный подход: ЕС создает не отдельные проекты, а целостную экосистему для генерации статистики на основе ДЗЗ (платформа EDP, методология, координация).

2. Акцент на валидации: Все результаты обязательно проверяются через наземные обследования (LUCAS) и административные данные (IACS), что обеспечивает высокое качество официальной статистики.

3. Политическая значимость: Использование ДЗЗ напрямую завязано на выполнение ключевых политик ЕС: Общей сельскохозяйственной политики (CAP) и Европейского зеленого курса (European Green Deal).

4. Открытость и прозрачность: Все исходные данные программы Copernicus находятся в свободном доступе, что стимулирует инновации и позволяет независимо проверять результаты.

Таким образом, ЕС демонстрирует наиболее зрелый и комплексный подход к интеграции данных ДЗЗ в официальную статистику, превращая их из экспериментального инструмента в один из основных источников данных для мониторинга социально-экономических и экологических процессов.

2. Нормативное правовое регулирование

Правовая база ЕС создает мандат и обязательства для использования новых источников данных, включая ДЗЗ.

На уровне ЕС:

1. Регламент (ЕС) 2019/2152 о европейской статистике бизнеса:

Ключевое положение: Прямо обязывает Евростат и НСС использовать альтернативные источники данных (включая спутниковые данные) для снижения нагрузки на респондентов и повышения качества статистики. Это главный правовой мандат.

2. Регламент (EU) 2023/138 о комплексной системе экономических счетов и экологической деятельности:

Ключевое положение: Требует от стран-членов предоставлять детальные данные по экосистемным счетам. Данные ДЗЗ являются основным источником для выполнения этих требований (например, для карт землепользования и растительного покрова).

3. Директива INSPIRE (Infrastructure for Spatial Information in Europe):

Ключевое положение: Обязывает государства-члены создавать совместимые геопространственные данные и сервисы. Это обеспечивает стандартизацию и возможность совместного использования пространственных данных, включая результаты обработки ДЗЗ, между всеми госорганами.

4. Политика открытых данных Copernicus:

Ключевое положение: Закрепляет принцип полного, свободного и открытого доступа к данным со спутников Sentinel. Это снимает правовые барьеры, связанные с интеллектуальной собственностью и лицензированием.

На национальном уровне:

Национальные законы о статистике: Как правило, дублируют и конкретизируют положения регламентов ЕС, давая НСС право запрашивать и использовать административные и альтернативные данные.

Законы о геопространственной информации: Реализуют директиву INSPIRE на национальном уровне, создавая правовую основу для обмена и использования геоданных.

3. Методология на уровне ЕСС/Евростата и на национальном уровне

ЕС создал целостную методологическую экосистему для обеспечения качества, сопоставимости и воспроизводимости результатов.

На уровне ЕСС/Евростата:

2. Специализированные методологические руководства и пилотные модули (Pilot Production Modules - PPMs):

Содержание: Конкретные, готовые к использованию методики для расчета определенных показателей. Например:

PPM по классификации землепользования на основе LUCAS и Sentinel-2.

PPM по оценке урожайности.

PPM по мониторингу городского sprawl (для ЦУР 11.3.1).

Статус: Эти модули включают код (часто на Python или R), документацию и учебные материалы. Они развертываются на платформе EDP.

3. Рамочный стандарт качества ESS (Quality Assurance Framework - QAF):

Содержание: Все методы, основанные на ДЗЗ, должны соответствовать этому строгому стандарту, включая критерии релевантности, точности, своевременности, доступности и сопоставимости.

На национальном уровне:

1. Национальные методологические руководства:

Содержание: НСС адаптируют общеевропейские руководства под национальную специфику (например, под набор основных сельскохозяйственных культур, особенности ландшафта).

Пример: ISTAT (Италия) разработал детальную методику интеграции ДЗЗ, IACS и сельскохозяйственной переписи.

2. Процедуры валидации:

Содержание: Каждое НСС разрабатывает и документирует внутренние процедуры проверки результатов, полученных на основе ДЗЗ. Ключевой элемент — перекрестная проверка с традиционными источниками: данными переписи, обследований и административных реестров (в первую очередь, IACS).

3. Постоянное обучение и рабочие группы:

Содержание: Создаются внутренние рабочие группы и центры компетенций. Сотрудники проходят обучение как в Евростате (например,

вебинары ESS), так и на внешних курсах по машинному обучению и обработке спутниковых снимков.

Главный принцип методологии ЕС: «Единообразие через сотрудничество». Евростат задает общие стандарты и предоставляет инструменты (PPM, EDP), а НСС адаптируют их, обеспечивая при этом высокий уровень качества через строгие процедуры валидации. Это позволяет получать сопоставимые на уровне всего Европейского Союза данные, сохраняя гибкость для учета национальных особенностей.

Роль ЭСКАТО в проектах и методологии по ДЗЗ

Разработка и адаптация методологических руководств

Практические руководства по применению ДЗЗ:

ЭСКАТО, совместно с Продовольственной и сельскохозяйственной организацией ООН (ФАО) и Глобальной платформой ООН, разрабатывает и продвигает пошаговые методики использования спутниковых данных для решения конкретных задач. Примеры:

Сельское хозяйство: Оценка посевных площадей, мониторинг засух с использованием индекса вегетации (NDVI).

Окружающая среда: Мониторинг обезлесения, деградации земель, качества воды.

Урбанизация: Картографирование роста городов и оценка плотности застройки.

Создание и поддержка региональной ИТ-инфраструктуры: "The Asia-Pacific POP Hub"

Это одна из самых значимых практических инициатив ЭСКАТО.

Суть проекта: ЭСКАТО является куратором и соразработчиком регионального "центра обработки данных" (The Asia-Pacific POP Hub – Partner-Operate-Provide) на базе глобальной UN Global Platform.

Функционал: Этот хаб предоставляет странам региона, не имеющим собственных суперкомпьютеров, бесплатный облачный доступ к:

Данным ДЗЗ (Sentinel, Landsat).

Инструментам анализа (готовые алгоритмы для классификации земного покрова, обнаружения судов и т.д.).

Вычислительным мощностям для их обработки.

Связь с Китайским хабом: ЭСКАТО активно сотрудничает с Глобальным хабом по ДЗЗ в Ханчжоу, используя его экспертизу и алгоритмы, и делая их доступными для других стран региона через свою платформу.

Реализация пилотных проектов и наращивание потенциала

Проекты с данными ДЗЗ:

Обучение сотрудников НСС использованию платформы Open Foris (Collect Earth) и SEPAL для проведения национальных оценок землепользования и лесного покрова.

Создание региональных карт и статистики по индикаторам ЦУР, связанным с земельными и водными ресурсами.

Обучение и обмен опытом: ЭСКАТО проводит региональные и субрегиональные тренинги, вебинары и стажировки, фокусируясь на "преодолении цифрового разрыва" внутри региона.

Ключевые публикации и руководства, разработанные при ведущей роли ЭСКАТО

1. «The Guide to Using Earth Observation and Geospatial Information for Official Statistics in Asia and the Pacific»: Практическое руководство, шаг за шагом объясняющее, как интегрировать ДЗЗ в производственный процесс НСС.

2. Серия методических документов по расчету индикаторов ЦУР с использованием больших данных: Например, по индикаторам 11.3.1 (урбанизация), 15.3.1 (деградация земель).

3. Отчеты и аналитические записки по правовым и институциональным рамкам доступа к данным частного сектора, в том числе к данным мобильных операторов.

4. Документы по созданию и управлению национальными система данных (National Statistical Data Ecosystems), где большие данные рассматриваются как ключевой компонент.

Итог: ЭСКАТО выступает в роли «регионального интегратора и катализатора». Она не дублирует глобальные инициативы, а делает их

реализуемыми на местном уровне, предоставляя странам Азиатско-Тихоокеанского региона необходимые методологии, ИТ-инфраструктуру, обучение и площадку для кооперации, что необходимо для успешного и устойчивого внедрения больших данных в официальную статистику.

Роль ФАО в проектах и методологии по ДЗЗ

ФАО является одним из мировых лидеров по интеграции данных ДЗЗ и геопространственной информации в практику официальной статистики, прежде всего, в области продовольственной безопасности, сельского хозяйства и устойчивого управления природными ресурсами.

Ключевые направления использования ДЗЗ в ФАО:

Сельскохозяйственная статистика: Оценка площадей под посевами, мониторинг состояния посевов, прогнозирование урожайности, идентификация типов землепользования.

Статистика природопользования и экологии: Мониторинг вырубки и деградации лесов, оценка состояния водных ресурсов, опустынивания, запасов углерода.

Рыболовство и аквакультура: Мониторинг судов (AIS), оценка состояния аквакультурных ферм.

Раннее предупреждение о продовольственных кризисах: Мониторинг засух, наводнений, других экстремальных явлений, влияющих на производство продовольствия.

Система мониторинга сельскохозяйственных стрессов (ASIS)

Официальное название (англ.): Agricultural Stress Index System (ASIS).

Реквизиты: Глобальная методика, разработанная Отделом климата и природных ресурсов ФАО. Постоянно действующая система с 2014 года.

Содержание: Методика использует спутниковые данные для выявления сельскохозяйственных районов, подверженных засухе, на глобальном, национальном и субнациональном уровнях.

Платформа «Hand-in-Hand» Geospatial Platform

Официальное название (англ.): FAO Hand-in-Hand Geospatial Platform.

Реквизиты: Запущена ФАО в 2020 году как часть инициативы «Hand-in-Hand» для ускорения преобразования агропродовольственных систем.

Содержание: Представляет собой обширный каталог тысяч геопространственных слоев (включая данные ДЗЗ и производные показатели), предоставляя НСС мощный инструмент для анализа и принятия решений.

Статус внедрения в НСС и лучшие практики

Статус внедрения: Внедрение находится на разных стадиях. Развитые страны (Евростат, USDA) активно используют ДЗЗ в режиме оперативного мониторинга. Многие развивающиеся страны при поддержке ФАО (например, в рамках проектов в Кении, Гамбии, Шри-Ланке) проводят пилотные переписи и внедряют методы ДЗЗ для стратификации и оценки площадей.

Лучшие практики:

Индия: Оперативная оценка посевных площадей риса и пшеницы с использованием радарных данных (RISAT) для министерства сельского хозяйства.

Бразилия: Мониторинг вырубки лесов в Амазонии в реальном времени с помощью системы DETER на основе данных ДЗЗ.

Используемая ФАО ИТ-инфраструктура

ФАО развивает облачную и открытую инфраструктуру для обработки больших геоданных:

SEPAL (System for Earth Observation Data Access, Processing and Analysis for Land Monitoring):

Описание: Облачная платформа с открытым кодом, предоставляющая НСС доступ к вычислительным мощностям, спутниковым данным (Landsat, Sentinel) и готовым алгоритмам для классификации и анализа без необходимости закупать дорогостоящее оборудование и ПО.

Open Foris (Collect Earth, Calc, etc.):

Описание: Набор открытых инструментов, среди которых ключевой для НСС – Collect Earth. Это инструмент для визуального сбора данных на основе

спутниковых снимков высокого разрешения, позволяющий проводить оценку земельного покрова и его изменений, верификацию классификаций.

Перспективы и стратегии развития ФАО

Интеграция с другими источниками больших данных: Комбинирование ДЗЗ с данными сотовых операторов, социальных сетей, IoT-устройств в агросекторе.

Развитие искусственного интеллекта и машинного обучения: Создание более точных и автоматизированных алгоритмов для классификации и прогнозирования.

Повышение частоты и детальности наблюдений: Использование группировок малых спутников (CubeSats), таких как Planet, для ежедневного мониторинга.

Стратегическая программа «50x2030»: В рамках этой инициативы по укреплению сельскохозяйственной статистики в 50 странах к 2030 году, данные ДЗЗ определены как ключевой компонент для улучшения дизайна обследований и расчета показателей.

Развитие платформы «Hand-in-Hand»: Дальнейшая интеграция в нее новых инструментов анализа в реальном времени и прогнозных моделей.

Проекты ФАО в странах БРИКС и СНГ

Страны БРИКС:

Бразилия: Является мировым лидером в использовании ДЗЗ для мониторинга лесов (INPE). ФАО выступает преимущественно как площадка для обмена опытом Бразилии с другими странами.

Индия: ФАО сотрудничает с индийскими агентствами в области использования ДЗЗ для оценки состояния лесов и в рамках инициативы «Hand-in-Hand».

Китай: Активное сотрудничество в рамках глобальных систем мониторинга, обмен данными ДЗЗ.

ЮАР: ФАО оказывала поддержку в рамках мониторинга продовольственной безопасности и засух.

Характеристика проектов ФАО в Казахстане, Кыргызстане и Узбекистане

Основной фокус проектов ФАО в странах Центральной Азии смещен в сторону решения проблем деградации земель, управления водными и пастбищными ресурсами в условиях изменения климата с использованием ДЗЗ как ключевого инструмента для получения объективных данных.

Казахстан

Ключевой проект: Внедрение Системы мониторинга сельскохозяйственных стрессов (ASIS) и оценка пастбищ.

Цель: Повышение устойчивости агропромышленного комплекса к засухам и совершенствование системы управления пастбищами.

Деятельность в части ДЗЗ:

Адаптация методологии ASIS: Казахстанский научно-исследовательский институт экономики агропромышленного комплекса и развития сельских территорий при поддержке ФАО провел работу по адаптации глобальной методологии ASIS к национальным условиям. Это позволяет на регулярной основе получать карты индекса стресса растительности (VHI) для мониторинга засух на пастбищных и пахотных землях.

Оценка состояния пастбищ: С использованием данных ДЗЗ (Landsat, Sentinel-2) и платформы SEPAL проводятся работы по картографированию продуктивности пастбищ, оценке нагрузки на пастбищные экосистемы и выявлению деградированных территорий.

Партнеры: Министерство сельского хозяйства РК, КазНИИ АПК и РСТ.

Результат: Методология внедрена и используется для подготовки аналитических записок и прогнозов. Является примером успешной передачи технологий от ФАО национальному институту.

Кыргызстан

Ключевой проект: Устойчивое управление природными ресурсами в условиях изменения климая. Внедрение инструментов ДЗЗ для оценки земель.

Цель: Создание национального потенциала для мониторинга деградации земель и обоснования планов по адаптации к изменению климата.

Деятельность в части ДЗЗ:

Обучение использованию Collect Earth: ФАО провела серию интенсивных национальных тренингов для специалистов из Минсельхоза, Агентства по охране окружающей среды и научных кругов. Участники обучены проводить оценку земельного покрова и его изменений с помощью Collect Earth на основе снимков высокого разрешения (Google Earth, Bing Maps).

Пилотные проекты по оценке деградации земель: В пилотных районах были проведены оценки, совмещающие полевые исследования и дешифрирование данных ДЗЗ. Это позволило верифицировать карты деградации и оценить масштабы проблемы.

Партнеры: Министерство сельского, водного хозяйства и развития регионов, Государственное агентство охраны окружающей среды и лесного хозяйства.

Результат: В стране создан пул сертифицированных национальных экспертов, способных самостоятельно проводить мониторинг земельных ресурсов с использованием современных геопространственных методов. Полученные данные используются для отчетности в рамках КБО ООН (например, по ЦУР 15.3.1 "Доля деградированных земель").

Узбекистан

Ключевой проект: Внедрение климатически оптимизированных подходов в сельское хозяйство и мониторинг засух.

Цель: Повышение продуктивности и устойчивости сельского хозяйства, в частности, в критически важном регионе — Приаралье.

Деятельность в части ДЗЗ:

Мониторинг засух с помощью ASIS: ФАО оказала поддержку Узгидромету в использовании системы ASIS для оперативного мониторинга засушливых явлений на территории страны.

Оценка потенциала пастбищ в Приаралье: Используя данные ДЗЗ, проводятся работы по оценке состояния и продуктивности пастбищ в зоне, подверженной опустыниванию. Это необходимо для разработки планов по устойчивому животноводству.

Использование платформы SEPAL: Национальные специалисты обучены работе в облачной платформе SEPAL для самостоятельного проведения

анализа многолетней динамики растительного покрова, классификации землепользования.

Партнеры: Узгидромет, Министерство сельского хозяйства РУ.

Результат: Узбекистан активно интегрирует данные ДЗЗ в свою практику. Информация, полученная с помощью инструментов ФАО, используется для подготовки государственных программ по развитию регионов, подверженных засухам и деградации земель.

Общий вывод по региону: Во всех трех странах ФАО действует по схожей стратегии: не подмена национальных органов, а передача технологий (SEPAL, Collect Earth) и методологий (ASIS). Ключевой акцент делается на обучении, чтобы национальные специалисты могли самостоятельно и на регулярной основе использовать данные ДЗЗ для принятия управленческих решений в сфере АПК и природопользования.

НСС де-факто не являлись ключевыми партнерами в указанных проектах ФАО в Казахстане и Узбекистане. Проекты были сфокусированы на прикладных задачах оперативного мониторинга (засухи, состояние пастбищ, деградация земель). За эти направления в странах СНГ традиционно отвечают не НСС, а отраслевые министерства (сельского хозяйства) и специализированные агентства (гидромет, экологии).

Целевая аудитория проектов ФАО: Помощь ФАО была направлена именно на те органы, которые принимают оперативные управленческие решения в сфере АПК и природопользования. Например, карты засух от Узгидромета нужны Минсельхозу для планирования мероприятий, а не НСС для расчета валового сбора.

Статус методологий ДЗЗ: На момент реализации проектов методы ДЗЗ еще не были в полной мере стандартизированы и легитимизированы для производства официальных статистических данных. Они рассматривались как ценный, но все же вспомогательный или экспериментальный инструмент.

Опосредованное влияние и перспективы интеграции:

1. Стратификация и улучшение рамок выборки: НСС могут использовать карты землепользования, созданные в рамках этих проектов (например, с помощью Collect Earth), для совершенствования основы выборки

сельскохозяйственных обследований. Это позволяет более точно определять целевые совокупности и снижать ошибку покрытия.

2. Верификация и оценка качества данных: Данные ДЗЗ могут использоваться НСС для перекрестной проверки (верификации) данных, поступающих от сельхозпроизводителей или из административных источников. Например, для подтверждения факта использования земельного участка.

3. Подготовка к будущим переписям: Как следует из Мировой программы переписи сельского хозяйства 2020 (WCA 2020), ДЗЗ рекомендованы для использования на всех этапах переписи. Опыт, накопленный отраслевыми ведомствами, создает национальный потенциал, которым НСС могут воспользоваться при проведении следующей сельскохозяйственной переписи. Они могут либо привлечь специалистов из этих ведомств, либо перенять методики.

4. Расширение показателей ЦУР: Для отчетности по Целям устойчивого развития (например, индикатор 15.3.1 по деградации земель) НСС часто являются координирующим органом. Данные, полученные экологическими и сельскохозяйственными ведомствами с помощью ДЗЗ, могут быть официально канализированы через НСС для представления в международные организации.

Общая система нормативного правового регулирования использования больших данных в официальной статистике ЕС

Общая Иерархия и Соотношение

1. Первичное право ЕС (Договоры): Высшая сила.

2. Регламенты (Regulations): Имеют прямое действие на всей территории ЕС и обязательны для всех стран-членов. Рассматриваемые в рамках настоящего Приложения акты - это в основном регламенты.

3. Директивы (Directives): Обязательны для достижения результата, но оставляют форму и методы на усмотрение национальных властей.

4. Национальное законодательство: Должно соответствовать праву ЕС. Конкретизирует и реализует директивы, а также может устанавливать дополнительные нормы в рамках, дозволенных регламентами.

Характеристика Ключевых Актов

1. GDPR (General Data Protection Regulation) - Регламент (ЕС) 2016/679

Статус: Рамочный ограничитель и основа легитимности. Задает строгие рамки для любой обработки персональных данных.

Ключевые положения для официальной статистики:

Статья 5: Принципы обработки (законность, честность, прозрачность, минимизация данных).

Статья 6: Правовые основания. Для статистики используется «выполнение задачи, осуществляемой в общественных интересах».

Статья 89: Позволяет обрабатывать персональные данные для статистических целей при условии применения принципов «минимализации данных» и «обезличивания». Это правовая основа для использования агрегированных и анонимизированных больших данных.

2. Data Governance Act (DGA) - Регламент (ЕС) 2022/868

Статус: Порядок доступа к данным. Создает механизмы для увеличения доверия к обмену данными и их повторному использованию.

Ключевые положения для официальной статистики:

Создает концепцию «посредников в обмене данными» (data intermediarie), которые участвуют в доверенном обмене данными, сохраняя нейтралитет.

Поощряет «альтруистический обмен данными» (data altruism), когда физические и юридические лица добровольно предоставляют свои данные для использования в общественных интересах, включая официальную статистику.

Упрощает повторное использование защищенных данных государственного сектора.

Значение для официальной статистики: Открывает новые каналы для получения данных от бизнеса и граждан на добровольной, но регулируемой основе.

3. Data Act - Регламент (ЕС) 2023/2854

Статус: Ключевой акт в отношении данных Интернета Вещей (IoT). Регулирует справедливый доступ и использование данных, генерируемых подключенными устройствами.

Ключевые положения для официальной статистики:

Статья 14 (Доступ к данным органами государственной власти): Прямо разрешает государственным органам, включая НСС, запрашивать и получать данные у владельцев данных (например, у производителей умной техники) при наличии «исключительной потребности» для выполнения задач в общественных интересах. При этом определение «исключительной потребности» включает «формирование официальной статистики».

Это инновационная норма, создающая прямое право НСС на доступ к альтернативным видам данных.

4. Регламент о европейской статистике № 223/2009 (с последующими поправками)

Статус: «Конституция» для Европейской статистической системы (ЕСС)". Определяет мандат, принципы и организацию.

Ключевые положения:

Транслирует Основополагающие принципы официальной статистики ООН на ЕС, устанавливает роль ЕСС, Евростата и НСС.

5. Регламент (ЕС) 2024/3018 (Поправка к Регламенту 223/2009)

Статус: Стимул для использования больших данных и модернизации ЕСС. Самая важная последняя новация.

Ключевые новации для больших данных:

Явный мандат на использование новых источников данных: Прямо обязывает Евростат и НСС «расширять использование новых источников данных и технологий» для повышения качества, снижения нагрузки на респондентов и повышения эффективности.

Упрощение доступа к административным данным: Упрощает процедуры для НСС по получению доступа к административным данным других госорганов.

Создание Единой Статистической Производственной Среды (ESPE): Закрепляет на законодательном уровне создание и развитие таких платформ, как EDP (Esper Big Data Platform), как ключевого инструмента ЕСС.

Основные последние новации и итоги

1. Смена парадигмы: От подхода «можно, если не запрещено» к подходу «нужно, и для этого созданы инструменты». Data Act и Поправка 2024/3018 — это мощные сигналы и инструменты для НСС.

2. Создание комплексной экосистемы: Законодательство ЕС больше не просто защищает данные (GDPR), но и активно создает механизмы для их легального использования в общественных интересах (DGA, Data Act).

3. Укрепление инфраструктуры: Закрепление ESPE/EDP в регламенте придает этим проектам устойчивость и долгосрочность, переводя их из разряда пилотов в основу будущего статистического производства.

4. Вызов для национального уровня: Теперь основное давление переносится на национальные правительства. Странам-членам ЕС необходимо:

- привести национальные законы о статистике в соответствие с новыми регламентами, предоставив НСС четкие мандаты.

- наладить механизмы реализации Data Act и DGA на национальном уровне, чтобы НСС могли реально воспользоваться этими правами.

Таким образом, ЕС создал, самую передовую в мире законодательную экосистему для интеграции больших данных в официальную статистику, сочетающую в себе строгую защиту прав граждан с инновационными инструментами доступа к данным для общественного блага.

Стратегии по данным в ЕС

1. **Общеевропейские стратегии (программы)**

Стратегия: «A European strategy for data» (COM(2020) 66 final)

Статус: Стратегическая коммуникация Европейской комиссии, задающая вектор до 2030 года. Является основой для последующих регламентов (Data Governance Act, Data Act).

Ключевые цели, релевантные для официальной статистики:

Создание «Единого рынка данных» (European Single Market for Data), где данные могут свободно перемещаться внутри ЕС при соблюдении правил.

Продвижение концепции «суверенитета данных» (Data Sovereignty) ЕС.

Создание секторальных «общих пространств данных» (Common European Data Spaces) в ключевых областях (здравоохранение, сельское хозяйство, транспорт и т.д.).

Значение для официальной статистики: НСС рассматриваются как одни из ключевых «потребителей» и «бенефициаров» этих пространств данных. Стратегия прямо создает экосистему, в которой статистическим службам будет существенно проще получать доступ к данным.

Программа: «Digital Decade Policy Programme 2030» (Decision (EU) 2022/2481)

Статус: Обязательная для исполнения программа ЕС до 2030 года.

Ключевые цели (цифровые компетенции):

К 2030 году 80% взрослого населения должны иметь базовые цифровые навыки.

20 млн специалистов по ИКТ в ЕС.

Ускорение цифровой трансформации бизнеса и госуслуг.

Значение для официальной статистики: Создает благоприятную среду: более цифровое общество и экономика генерируют больше «цифровых следов» (больших данных), а также готовит кадры, необходимые НСС для работы с этими данными.

Позиция: «ESS Common Position on the future and strategic priorities of European statistics» (21.05.2025)

Статус: Совместная стратегическая декларация Евростата и всех глав НСС стран-членов ЕС – «дорожная карта» развития ЕСС.

Ключевые приоритеты:

1. Полная интеграция новых источников данных: План по переходу от пилотов к регулярному использованию больших данных.

2. Внедрение ИИ: Использование искусственного интеллекта для автоматизации и улучшения анализа.

3. Совершенствование инфраструктуры: Развитие ESPE/EDP и доверенных сред.

4. Развитие кадрового потенциала: Переобучение статистиков в специалистов (ученых) по данным.

5. Укрепление партнерств: С частным сектором, наукой, гражданским обществом.

2. Национальные стратегии по данным НСС (примеры передовой практики)

Нидерланды: Statistics Netherlands (CBS) - «Strategy 2020-2025: Data for a better future»

Принята: Statistics Netherlands, 2020 г.

Период действия: 2020-2025 гг.

Цели и задачи:

Уход от традиционных обследований: Максимально использовать административные и большие данные.

Создание «Цифрового двойника» общества: Комплексное моделирование социально-экономических процессов.

Укрепление роли НСС как «Национального Координатора Данных»: Агрегация данных из разных источников для госорганов.

Италия: ISTAT - «Strategic Plan 2021-2017» (Piano Strategico 2021-2027)

Принят: ISTAT, 2021 г.

Период действия: 2021-2027 гг.

Цели и задачи:

Инновации в данных (Dati+): Систематическое использование больших данных, в том числе сотовых операторов, ДЗЗ, веб-данных.

Расширение аналитики (Analisi+): Внедрение методов ИИ и ML.

Улучшение распространения (Diffusione+): Интерактивные дашборды и API.

Норвегия: Statistics Norway (SSB) - «Strategy for Data-Driven Development 2021-2025»

Принята: Statistics Norway, 2021 г.

Период действия: 2021-2025 гг.

Цели и задачи:

Данные как основа всей деятельности: Принцип «data first».

Расширенное использование административных данных и начало использования больших данных.

Методы ИИ для автоматизации и анализа.

Финляндия: Statistics Finland - «Strategy 2025»

Принята: Statistics Finland, 2021 г.

Период действия: до 2025 г.

Цели и задачи:

Своевременная и детальная статистика на основе новых источников данных.

Снижение нагрузки на респондентов.

Активное использование ИИ.

Великобритания: UK Statistics Authority - «Statistics for the Future: The UK Statistics Authority's Strategy 2020-2025»

Принята: UK Statistics Authority, 2020 г.

Период действия: 2020-2025 гг.

Цели и задачи:

Максимизация использования новых источников данных: Партнерства для доступа к новым данным.

Внедрение передовых методов: Включая науку о данных и ИИ.

Создание более полной картины общества и экономики.

Общий вывод: На общеевропейском и на национальном уровне произошел стратегический сдвиг. Большие данные и инновации перестали быть периферийной активностью и стали центральным элементом стратегий развития официальной статистики в цифровую эпоху. Все стратегические документы нацелены на системную трансформацию, а не на отдельные проекты.

Центры компетенций по большим данным в ЕС

Общеевропейские центры компетенций и инфраструктура

1. Platform for Big Data in Official Statistics (EDP - Esper Big Data Platform)

Облачная платформа, физически развернута в дата-центрах ЕС.

Создана: Запуск в 2020 году.

Учредители: Евростат в тесном сотрудничестве с национальными статистическими институтами (НСС) ЕС.

Партнеры: Все НСС стран-членов ЕС, Европейское космическое агентство (ЕКА), коммерческие провайдеры облачных услуг.

Цели и задачи:

Предоставить единую, безопасную и масштабируемую облачную среду для обработки больших данных всеми статистическими службами ЕСС.

Устранить необходимость для каждого НСС создавать собственную дорогостоящую ИТ-инфраструктуру.

Стандартизировать рабочие процессы (pipelines) и обеспечить воспроизводимость результатов.

Основные проекты:

Пилотные производственные модули (Pilot Production Modules - PPMs):

Готовые алгоритмы обработки данных, например:

PPM LUCAS: Классификация землепользования по данным ДЗЗ и наземных обследований.

PPM на основе данных СО: Анализ мобильности и туризма.

Интеграция с данными Copernicus: Прямой доступ и обработка спутниковых снимков Sentinel-1, Sentinel-2 и др.

2. ESS Labs Network (Сеть лабораторий ECC)

Виртуальная сеть, объединяющая физические лаборатории при НСС.

Создана: Инициатива запущена около 2018 года.

Учредители: Евростат.

Партнеры: Более 20 национальных «лабораторий» больших данных, включая ISTAT (Италия), CBS (Нидерланды), INE (Португалия), SSB (Норвегия), Destatis (Германия).

Цели и задачи:

Создать сообщество практиков по большим данным в официальной статистике.

Быстро тестировать и валидировать новые источники данных и методики.

Обмениваться кодом, ноу-хау и лучшими практиками.

Основные проекты: Совместные пилотные проекты по использованию данных СО, веб-скрепинга, спутниковых снимков, данных AIS (судоходство).

Ключевые национальные центры компетенций (Хабы)

1. Италия: ISTAT - Лаборатория больших данных (Big Data Lab)

Местоположение: Рим, Италия.

Создана: Одна из первых и наиболее активных в ECC, создана в 2015-2016 гг.

Учредители: Национальный институт статистики Италии (ISTAT).

Партнеры: Университеты, исследовательские центры, операторы сотовой связи, компании-агрегаторы данных.

Цели и задачи: Разработка и внедрение методов использования больших данных для производства официальной статистики, в первую очередь, в сфере туризма, мобильности населения, сельского хозяйства.

Основные проекты:

Мобильность и туризм на основе данных СО: Лидер в ЕС по производству регулярной статистики туристских потоков.

Цены жилья на основе веб-данных: Анализ объявлений о продаже/аренде недвижимости.

Сельскохозяйственная статистика на основе данных ДЗЗ.

Нидерланды: Statistics Netherlands (CBS) - Центр Больших Данных (Center for Big Data Statistics - CBDS)

Местоположение: Гаага/Херлен, Нидерланды.

Создан: 2014 год.

Учредители: Statistics Netherlands (CBS).

Партнеры: Более 30 организаций, включая IBM, Microsoft, Oracle, университеты Твенте, Делфта, Амстердама, а также другие НСС.

Цели и задачи: Исследовать, разрабатывать и внедрять инновационные методы производства официальной статистики на основе больших данных.

Основные проекты:

CBS Mobility Index: Измерение мобильности на основе данных СО и GPS-данных.

Прогнозирование экономических показателей с использованием данных о трафике с грузовиков, потреблении энергии и др.

Мониторинг ЦУР с помощью ДЗЗ и других данных.

Португалия: Statistics Portugal (INE) - Лаборатория Больших Данных и Искусственного Интеллекта (Big Data & AI Lab)

Местоположение: Лиссабон, Португалия.

Создана: 2021 год.

Учредители: Statistics Portugal (INE).

Партнеры: Португальское Агентство по туризму, операторы мобильной связи, академические учреждения.

Цели и задачи: Сосредоточена на использовании больших данных и ИИ для модернизации статистики, в первую очередь, в области туризма, мобильности населения и региональной статистики.

Основные проекты:

Статистика туризма на основе данных СО: Производство ежемесячных данных о туристах.

Проект AURORAL: (в рамках Horizon 2020) по созданию платформы для обмена данными в сельских районах, с использованием ДЗЗ и IoT.

Норвегия: Statistics Norway (SSB) - Отдел больших данных и интеграции данных

Местоположение: Осло/Конгсвингер, Норвегия.

Создан: Структурно оформился в 2010-х годах.

Учредители: Statistics Norway (SSB).

Партнеры: Норвежские государственные органы, университеты, коммерческие компании.

Цели и задачи: Использование больших данных и административных регистров для повышения качества и эффективности официальной статистики.

Основные проекты:

Статистика туризма на основе данных СО.

Использование данных AIS для статистики морских перевозок и рыболовства.

Интеграция данных ДЗЗ для мониторинга землепользования и лесов.

«Песочницы» (Sandboxes) и доверенные среды

Концепция «песочницы» в ЕСС часто реализуется через саму платформу EDP и доверенные среды обработки (Trusted Smart Satistics - TSS) при национальных институтах.

Концепция TSS: Это безопасная изолированная среда, в которую поставщики данных (например, СО) загружают свои данные. Статистики

получают доступ к этой среде для кодирования и анализа, но не могут выгрузить исходные данные, обеспечивая их конфиденциальность.

Пример: Проект TSS multi-MNO (GOPA): Этот пилот был реализован именно в формате TSS-песочницы, в которую несколько сотовых операторов предоставили свои данные для совместного анализа.

Общий вывод: ЕС создал целостную, многоуровневую инновационную экосистему, которая включает в себя:

1. Централизованную инфраструктуру (EDP).

2. Сеть центров компетенций и сотрудничества (ESS Labs).

3. Мощные национальные ИТ хабы (в CBS, ISTAT и др.), которые являются двигателями инноваций.

4. Правовые и методологические рамки, обеспечивающие качество и безопасность.

Эта экосистема позволяет ЕСС системно и скоординированно двигаться к цели полной интеграции новых источников данных в официальную статистику.

Оценка качества официальных статистических данных при использовании больших данных

Традиционная основа:

Евростат: «Quality Assurance Framework of the European Statistical System (ESS QAF)», версия 2.0 (2022 г.).

ООН: «National Quality Assurance Frameworks (NQAF) Manual», версия 2.0 (2019 г.).

Ключевое изменение: Эти рамочные руководства были созданы для данных, формируемых в контролируемых процессах (обследования, регистры).

Характеристика специфики оценки по компонентам качества

Трансформация основных классических компонент качества в контексте больших данных:

1. Точность и надежность (Accuracy and Reliability)

Специфика больших данных:

Смещение (Bias): Из случайной ошибки превращается в систематическое смещение. Данные СО не включают детей и стариков, данные соцсетей смещены в сторону молодежи. Задача — не устранить, а измерить и скорректировать это смещение.

Погрешность: Источником становится не только выборка, но и алгоритм обработки. Точность модели классификации (F1-score) становится новой мерой статистической погрешности.

Требования руководств: ESS QAF 2.0 (Раздел 3.3) требует оценки и документирования всех источников погрешности. Для больших данных это означает обязательное вычисление и публикацию метрик качества алгоритмов (precision, recall) и оценок смещения.

Своевременность и пунктуальность (Timeliness and Punctuality)

Специфика больших данных: Главное преимущество. Позволяет перейти к near-real-time статистике. Однако возникает компромисс «timeliness vs. accuracy»: первая, быстрая оценка может быть менее точной.

Требования руководств: NQAF 2.0 (п. 2.4.1) требует оптимизации временного лага. Для больших данных это означает создание многоуровневой системы распространения: сначала быстрые оперативные индикаторы, затем проверенные и откалиброванные данные.

Сопоставимость (Comparability)

Специфика больших данных: Частые «разломы сопоставимости» при смене алгоритма или источника данных. Обновление модели ML может сделать временные ряды несопоставимыми.

Требования руководств: ESS QAF 2.0 (Раздел 3.5) требует предупреждения пользователей о разрывах в рядах. Для больших данных это означает ведение параллельных рядов при обновлении методологии и четкое версионирование алгоритмов.

Когерентность (Coherence)

Специфика больших данных: Когерентность не достигается сама собой, а является результатом сложной интеграции. Данные СО о населении должны быть согласованы с данными переписи и регистров.

Требования руководств: NQAF 2.0 (п. 2.6.1) требует обеспечения согласованности между различными статистическими продуктами. Для больших данных это требует разработки методов калибровки и статистического взвешивания для согласования с традиционными источниками.

2. Новые критические компоненты качества больших данных

Помимо адаптации старых, появляются новые компоненты, без которых обеспечение качества невозможно:

Надежность источника и прозрачность методологии (Source Credibility & Methodological Transparency)

Описание: Откуда данные? Почему мы доверяем оператору связи или платформе соцсетей? Как именно алгоритм ML пришел к результату?

Отражение в руководствах: ESS QAF 2.0 прямо указывает на необходимость прозрачности методов и управления метаданными. Это требует «Методологических паспортов», описывающих весь конвейер данных, а не только финальные результаты.

Воспроизводимость и стабильность алгоритма (Reproducibility & Algorithmic Stability)

Описание: Приведет ли повторный запуск кода с теми же данными к тем же результатам? Не «сойдет ли алгоритм с ума» при поступлении новых данных?

Отражение в руководствах: Это прямое требование принципа научности основополагающих принципов официальной статистики ООН. Реализуется через версионирование кода, контейнеризацию (Docker) и сквозное документирование всех параметров модели.

Этичность и конфиденциальность (Ethics & Privacy)

Описание: Качество статистики, полученной с нарушением этических норм и прав граждан, является псевдокачеством. Конфиденциальность — неотъемлемая часть качества.

Отражение в руководствах: NQAF 2.0 (Принцип 8) и ESS QAF 2.0 содержат строгие требования по защите данных. Для больших данных это означает использование дифференциальной приватности, синтетических данных и методов безопасной аналитики (Secure Multi-Party Computation).

Итог: Использование больших данных не отменяет существующие оценки качества, но требует их серьезного углубления. НСС должны научиться измерять и документировать не только погрешность измерения, но и погрешность алгоритма, смещение источника и стабильность IT-конвейера. Качество в эпоху больших данных — это качество не только результата, но и всего процесса, от источника до алгоритма.

Системы мониторинга и контроля качества в НСС стран СНГ

Все страны СНГ в той или иной форме используют системы контроля качества, основанные на рамочных принципах ООН. Однако указанные системы контроля качества в настоящее время не учитывают специфики формирования официальной статистической информации с использованием больших данных.

Международные стандарты в области больших данных

Перечень и характеристика стандартов ISO в области больших данных

Работу по стандартизации в этой области ведет подкомитет SC 42 «Искусственный интеллект» в составе технического комитета ISO/IEC JTC 1 «Информационные технологии».

Фундаментальный стандарт: Серия ISO/IEC 20546 — Общий обзор и словарь

Официальное название на английском: ISO/IEC 20546:2019 Information technology — Big data — Overview and vocabulary.

Краткая характеристика:

Данный стандарт является основополагающим. Он предоставляет общепризнанное понимание больших данных на международном уровне.

Основная функция: Дает согласованные определения ключевых терминов и концепций.

Содержание: Включает обзор экосистемы больших данных, описывает их характеристики (расширяя классическую "3V" до более комплексных моделей), содержит подробный глоссарий.

Значение для официальной статистики: Позволяет НСС использовать единую, признанную в мире терминологию в своих методологических документах, стратегиях и при взаимодействии с партнерами.

Рамочный стандарт: Серия ISO/IEC 20547 — Эталонная модель больших данных

Серия ISO/IEC 20547. Полная структура:

Серия: ISO/IEC 20547 Информационные технологии — Эталонная модель больших данных (Big data reference architecture — BDRA)

Часть 1: ISO/IEC 20547-1:2020 — Framework and process for reference architecture (Структура и процесс для эталонной архитектуры)

Характеристика: Задает общие принципы и методологию разработки эталонной модели.

Часть 2: ISO/IEC 20547-2:2020 — Use cases and derived requirements (Варианты использования и производные требования)

Характеристика: Содержит каталог типовых сценариев (use cases) применения больших данных. На основе этих сценариев выводятся функциональные и нефункциональные требования к эталонной архитектуре, описанной в Части 3.

Часть 3: ISO/IEC 20547-3:2020 — Reference architecture (Эталонная архитектура)

Характеристика: Ядро серии, описывающее концептуальную модель и логические компоненты экосистемы больших данных (как было указано ранее).

Часть 4: ISO/IEC 20547-4:2023 — Security and privacy protection (Защита безопасности и конфиденциальности)

Характеристика: Детализирует сквозную функцию «Security and Privacy», обозначенную в Части 3. Описывает риски, уязвимости и меры контроля для обеспечения безопасности и конфиденциальности на всех этапах работы с большими данными.

Часть 5: ISO/IEC 20547-5:2023 — Standards roadmap (Дорожная карта стандартов)

Характеристика: Этот документ предоставляет обзор существующих и разрабатываемых стандартов, имеющих отношение к экосистеме больших данных. Помогает организациям ориентироваться в ландшафте стандартизации и планировать их внедрение.

Таким образом, серия ISO/IEC 20547 представляет собой законченный и комплексный пакет: от методологии и требований (Ч.1, Ч.2) через архитектурное ядро (Ч.3) к критически важным аспектам безопасности (Ч.4) и навигации по стандартам (Ч.5).

Стандарты, касающиеся анализа данных

Серия ISO/IEC 24668: Процессы анализа больших данных

Официальное название на английском: ISO/IEC 24668:2022 Information technology — Big data — Process of data analytics

Краткая характеристика:

Стандартизирует жизненный цикл и основные процессы анализа больших данных.

Основная функция: Описывает этапы работы с данными: от постановки задачи и сбора данных до их очистки, анализа, интерпретации результатов и документирования.

Значение для официальной статистики: Позволяет формализовать и стандартизировать внутренние процедуры анализа, что повышает воспроизводимость, качество и прозрачность статистических продуктов, созданных на основе больших данных.

Стандарты по большим данным в ЕС и странах СНГ

Европейский Союз (ЕС)

В ЕС не создаются отдельные, автономные стандарты на большие данные, аналогичные ISO. Вместо этого используется стратегия ссылки на стандарты (referencing), в первую очередь, разработанные международными организациями по стандартизации (ISO/IEC, ITU-T).

Ключевой механизм: Регламент (ЕС) 1025/2012 Европейского парламента и Совета о европейской стандартизации.

Официальное название на английском: Regulation (EU) No 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardization.

В рамках этого регламента Европейская комиссия может выдавать «задания на стандартизацию» (Standardisation Request) признанным европейским организациям по стандартизации (CEN, CENELEC, ETSI).

Текущий статус по большим данным:

Прямого европейского стандарта (EN), дублирующего или заменяющего серию ISO/IEC 20547, на данный момент не существует.

Стандарты ISO/IEC де-факто являются основными для использования в проектах ЕС. Например, архитектура больших данных, продвигаемая в рамках программы Digital Europe, концептуально полностью соответствует ISO/IEC 20547-3.

Акцент в ЕС смещен в сторону регуляторных актов, устанавливающих правила для данных как таковых, а не их архитектурной обработки.

Страны СНГ

В странах СНГ ситуация неоднородна, но в целом находится на начальной стадии.

В рамках Межгосударственного совета по стандартизации (МГС) страны СНГ работают над принятием единых стандартов, часто основанных на адаптации ISO. Национальные институты стандартизации стран СНГ, как правило, также ориентируются на прямые переводы и внедрение стандартов ISO/IEC.

Основное внимание уделяется не столько архитектурным стандартам, сколько техническим регламентам и законам, касающимся защиты персональных данных и информационной безопасности, что создает правовой контекст для работы с большими данными.

Стандарты по большим данным в Республике Беларусь

В Беларуси был разработан и введен в действие собственный стандарт.

Официальное название на русском: СТБ ИСО/МЭК 20546-2020 «Информационные технологии. Большие данные. Обзор и словарь»

Реквизиты: Дата введения: 2021-04-01. Является полным аналогом международного стандарта ISO/IEC 20546:2019.

Краткая характеристика: Как и его международный и российский аналоги, этот стандарт устанавливает основополагающие термины и определения, что создает единую терминологическую базу для дальнейшего развития направления в стране.

Стандарты по большим данным в России

В России работа ведется в рамках Технического комитета по стандартизации ТК 001 «Искусственный интеллект» при Росстандарте. Помимо адаптации стандартов ISO, разрабатываются также собственные стандарты.

Принятые стандарты (адаптированные международные)

1. ГОСТ Р ИСО/МЭК 20546-2021 «Информационные технологии. Большие данные. Обзор и словарь»

2. Серия ГОСТ Р, аутентичная серии ISO/IEC 20547 (Эталонная модель больших данных)

2.1. Часть 1: Фундамент и процессы

Официальное название: ГОСТ Р ИСО/МЭК 20547-1-2023 «Информационные технологии. Эталонная модель больших данных. Часть 1. Структура и процесс для эталонной архитектуры»

Реквизиты: Дата введения: 2024-07-01. Заменяет собой предварительный стандарт ПНСТ 468-2021.

2.2. Часть 2: Варианты использования

Официальное название: ГОСТ Р ИСО/МЭК 20547-2-2023 «Информационные технологии. Эталонная модель больших данных. Часть 2. Варианты использования и производные требования»

Реквизиты: Дата введения: 2024-07-01.

2.3. Часть 3: Эталонная архитектура (ядро)

Официальное название: ГОСТ Р 56888-2023/ISO/IEC 20547-3:2020 «Информационные технологии. Эталонная модель больших данных. Часть 3. Эталонная архитектура»

Реквизиты: Дата введения: 2024-01-01.

2.4. Часть 4: Безопасность и конфиденциальность

Официальное название: ГОСТ Р ИСО/МЭК 20547-4-2023 «Информационные технологии. Эталонная модель больших данных. Часть 4. Безопасность и защита конфиденциальности»

Реквизиты: Дата введения: 2024-07-01.

2.5. Часть 5: Дорожная карта стандартов

Официальное название: ГОСТ Р ИСО/МЭК 20547-5-2023 «Информационные технологии. Эталонная модель больших данных. Часть 5. Дорожная карта стандартов»

Реквизиты: Дата введения: 2024-07-01.

3. Другие ключевые принятые стандарты

Стандарт по процессу аналитики данных (аутентичный международному)

Официальное название: ГОСТ Р ИСО/МЭК 24668-2023 «Информационные технологии. Большие данные. Процесс анализа данных»

Реквизиты: Дата введения: 2024-07-01.

Стандарт на техническое задание (оригинальный российский)

Официальное название: ГОСТ Р 57799-2023 «Информационные технологии. Большие данные. Техническое задание на создание системы больших данных»

Реквизиты: Дата введения: 2024-07-01.

Характеристика: Это важный оригинальный российский стандарт, аналога которому в линейке ISO на данный момент нет. Он устанавливает требования к структуре и содержанию технического задания (ТЗ) на создание систем больших данных. Стандарт ориентирован на практическое применение.

Проекты стандартов (находятся в разработке)

Помимо принятых, ряд проектов национальных стандартов (ПНСТ) находятся на стадии обсуждения и утверждения:

ПНСТ 714-2023/ISO/IEC 20547-2:2020 «…Эталонная модель больших данных. Часть 2. Варианты использования и производные требования» (Проект аутентичного перевода).

ПНСТ 715-2023/ISO/IEC 20547-4:2023 «...Эталонная модель больших данных. Часть 4. Безопасность и защита конфиденциальности» (Проект аутентичного перевода).

ПНСТ 716-2023/ISO/IEC 24668:2022 «...Большие данные. Процесс анализа данных» (Проект аутентичного перевода).

Данный пакет стандартов является одним из самых передовых и полных в мире и создает серьезную основу для реализации проектов НСС по использованию больших данных.

Типовая модель производства статистической информации ООН (ТМПСИ) версии 5.1 и 5.2

Характеристика GSBPM версии 5.1

Полное наименование (англ.): Generic Statistical Business Process Model (GSBPM), Version 5.1

Организация: Статистический отдел ООН (UNSD) в сотрудничестве с ЕЭК ООН (UNECE) и другими международными организациями.

Дата публикации: Январь 2019 года.

Статус: до настоящего времени являлась актуальной и широко принятой версией, служащей отраслевым стандартом для описания и реинжиниринга статистических бизнес-процессов.

Ключевые характеристики ТМПСИ 5.1:

1. Назначение: Модель предназначена для универсального описания полного жизненного цикла производства официальной статистики, независимо от источника данных (опросы, административные данные, большие данные).

2. **Структура: Модель иерархична и состоит из:**

8 основных этапов (Phases) высокого уровня.

45 подэтапов (Sub-processes), которые детализируют каждый этап.

Перекрестные процессы (Quality Management, Metadata Management, etc.), которые пронизывают все этапы.

Детализация этапов и подэтапов GSBPM v5.1:

Этап 1. Определение потребностей (Specify Needs): Определение потребностей пользователей и требований к статистическому продукту.

Подэтапы: Определение требований к выходным данным, проверка и приоритезация потребностей, установление производственных целей.

Этап 2. Проектирование (Design): Разработка всех аспектов дизайна статистического продукта.

Подэтапы: Проектирование выхода, проектирование сбора данных, проектирование обработки и анализа данных, проектирование метаданных, проектирование производственного процесса, проектирование архитектуры и ИТ, проектирование статистической методологии, оценка дизайна.

Этап 3. Сбор (Build): Создание и тестирование всех компонентов, необходимых для запуска производственного процесса.

Подэтапы: Построение/сборка инструментов сбора, построение ИТ-компонентов, конфигурация рабочих процессов, тестирование производственного процесса, тестирование статистической бизнес-архитектуры.

Этап 4. Сбор данных (Collect): Получение исходных данных из всех необходимых источников.

Подэтапы: Создание рамки и выборки, набор респондентов, запуск сбора данных, мониторинг сбора данных.

Этап 5. Обработка (Process): Преобразование собранных исходных данных в отдельные наборы данных, готовые для анализа.

Подэтапы: Интеграция данных, классификация и кодирование, редактирование и импутация, верификация и деривация, взвешивание и оценка, расчет агрегатов.

Этап 6. Анализ (Analyse): Подготовка статистических продуктов для распространения.

Подэтапы: Подготовка прототипов выходных данных, валидация выходных данных, интерпретация выходных данных и генерация пояснений.

Этап 7. Распространение (Disseminate): Выпуск статистических продуктов для пользователей.

Подэтапы: Управление выпуском, производство продуктов распространения, управление распространением, продвижение статистических продуктов.

Этап 8. Оценка (Evaluate): Общая оценка производственного цикла и продуктов.

Подэтапы: Сбор отзывов пользователей, оценка производственного процесса, оценка статистической бизнес-архитектуры.

Характеристика изменений в ТМПСИ версии 5.2

Полное наименование документа (англ.): Generic Statistical Business Process Model (GSBPM), Version 5.2

Организация: Статистический отдел ООН (UNSD).

Дата публикации: Май 2024 года.

Ключевой драйвер изменений: Адаптация модели к реалиям современного статистического производства.

1. Глобальные и кросс-процессуальные изменения

Изменение названия Этапа 4: С «Collect» (Сбор) на «Acquire» (Получение).

Обоснование: Термин «Collect» исторически ассоциировался с активным сбором данных через опросы. Новый термин «Acquire» более точно отражает процессы получения данных из разнообразных источников: загрузка административных данных, подключение к API, сбор потоков больших данных, скачивание спутниковых снимков и т.д.

Акцент на управлении данными (Data Management): Усилена интеграция принципов управления данными как сквозного процесса. Это отражает растущую важность таких аспектов, как управление метаданными, качество данных на уровне источника, управление мастер-данными и справочниками.

2. Изменения по этапам

Этап 1. Specify Needs (Определение потребностей)

Изменения: Уточнены описания подэтапов, чтобы подчеркнуть необходимость оценки пригодности и доступности новых источников данных на самом раннем этапе проектирования статистического продукта.

Этап 2. Design (Проектирование)

Подэтап 2.2: «Design data acquisition» (Проектирование получения данных) – новый подэтап, заменивший «Design data collection».

Содержание: Включает проектирование методов получения данных из всех типов источников: определение API, форматов файлов, прав доступа, методов веб-сканирования (web-scraping), интеграции с платформами ДЗЗ.

Подэтап 2.3: «Design data processing and staistical methods» (Проектирование обработки данных и статистических методов) – Объединенный и переименованный подэтап.

Содержание: Объединяет ранее разделенные процессы проектирования обработки и методологии, что отражает их тесную взаимосвязь, особенно при работе с большими данными, где алгоритмы обработки (например, машинное обучение) являются неотъемлемой частью методологии.

Этап 3. Build (Сборка)

Подэтап 3.1: «Build data acquisition instruments and components» (Сборка инструментов и компонентов получения данных) – Переименован.

Содержание: Включает не только создание электронных вопросников, но и разработку скриптов для API, ETL-процедур, конвейеров (pipelines) для потоковых данных.

Этап 4. Acquire (Получение) – Ключевое изменение

Подэтап 4.1: «Create frame and select sample» (Создание основы и выборки) – Расширен.

Содержание: Теперь явно включает создание основ для «неопросных» данных (например, списков веб-сайтов для сканирования, географических сеток для ДЗЗ).

Подэтап 4.2: «Set up acquisition» (Настройка получения) – новый подэтап.

Содержание: Замена «Recruit respondents». Включает настройку подключений к источникам, получение учетных данных, планирование автоматических загрузок.

Подэтап 4.3: «Run acquisition» (Запуск получения) – новый подэтап.

Содержание: Замена «Run collection». Непосредственный запуск процессов получения данных: выполнение ETL-заданий, активацию потоков данных, запуск сканирования.

Подэтап 4.4: «Finalise acquisition» (Завершение получения) – новый подэтап.

Содержание: Замена «Finalise collection». Включает подтверждение успешности получения, обработку сбоев, архивацию исходных данных.

Этап 5. Process (Обработка)

Изменения: Усилен акцент на автоматизированной и воспроизводимой обработке. В описаниях подэтапов появились отсылки к использованию методов машинного обучения для редактирования и импутации, а также к более сложным методам верификации и деривации, требуемым для больших данных.

Этап 6. Analyse (Анализ)

Изменения: Подчеркнута важность анализа и объяснения качества данных, полученных из новых источников, включая оценку смещений и репрезентативности.

Этапы 7 (Disseminate) и 8 (Evaluate)

Изменения: Не претерпели существенных структурных изменений, но их описания были обновлены для согласования с новой терминологией и акцентами всей модели.

Заключение: ТМПСИ 5.2 отражает сдвиг парадигмы в статистическом производстве и предоставляет НСС актуальный концептуальный каркас для трансформации своих процессов в сторону большей гибкости, автоматизации и эффективности при работе с современными источниками данных.

Специфика использования данных СО в рамках ТМПСИ v5.2

Ключевая особенность: Данные СО (CDR, xDR, локационные пинг-сигналы) являются пассивно-генерируемыми, неструктурированными для статистических целей и обладающими высокой скоростью поступления. Это коренным образом меняет подход к производству статистики по сравнению с традиционными опросами.

Этап 1. Определение потребностей (Specify Needs)

1.1. Определение требований к выходным данным: Требуется переформулировка потребностей пользователей в контексте новых возможностей. Наряду с «численностью населения на 1 января» возникает запрос на «ежедневную динамику численности и перемещений населения между регионами».

1.2. Проверка и приоритезация потребностей: Резко возрастает важность проверки этико-правовой осуществимости проекта. Потребность в данных о перемещениях должна быть сбалансирована с рисками для конфиденциальности.

1.3. Установление производственных целей: Цели смещаются от точечных измерений к непрерывному мониторингу. Например, цель: «Обеспечивать еженедельные показатели мобильности населения для нужд МЧС и градостроительства».

Этап 2. Проектирование (Design)

2.1. Проектирование выхода: Проектируются не статические таблицы, а интерактивные дашборды, API для распространения, динамические карты. Определяются новые показатели: «индекс мобильности», «коэффициент притока/оттока».

2.2. Проектирование получения данных (Acquisition): Ключевой подэтап. Проектируется конвейер (pipeline) данных: частота выгрузки CDR из систем оператора (ежедневно/еженедельно), методы безопасной передачи (шифрование, выделенные каналы), форматы данных (Avro, Parquet).

2.3. Проектирование обработки данных и статистических методов:

Методология: Разрабатываются статистические модели и алгоритмы для преобразования сырых сетевых событий в статистические показатели. Например, алгоритм «home-work detection» для определения места жительства и работы, модели дедубликации записей одного абонента.

Обработка: Проектируется сложная цепочка: анонимизация, агрегация, фильтрация «шумных» данных, привязка вышек сотовой связи к географическим полигонам.

2.5. Проектирование метаданных: Создается расширенный набор метаданных: не только о выходном показателе, но и о качестве и параметрах

исходных больших данных (коэффициент прореживания, доля анонимизированных записей, используемая версия алгоритма).

Этап 3. Сборка (Build)

3.1. Сборка инструментов и компонентов получения данных: Разрабатываются не анкеты, а скрипты для ETL/ELT-процессов, коннекторы к API оператора, решения для потоковой обработки (например, на базе Apache Kafka/Spark).

3.2. Сборка ИТ-компонентов: Создается высокомасштабируемая ИТ-инфраструктура: облачные хранилища (Object Storage), вычислительные кластеры для обработки петабайтов данных, системы управления контейнерами (Docker, Kubernetes).

3.4. Тестирование производственного процесса: Проводится нагрузочное тестирование всего конвейера на объемах данных, превышающих ожидаемые. Тестируется отказоустойчивость.

Этап 4. Получение (Acquire)

4.2. Настройка получения (Set up acquisition): Вместо подготовки интервьюеров осуществляется настройка автоматических задач (cron jobs) для запуска процессов выгрузки и передачи данных от оператора.

4.3. Запуск получения (Run acquisition): Процесс полностью автоматизирован. Данные начинают поступать в статистическое ведомство по заранее настроенным каналам без вмешательства человека.

4.4. Завершение получения (Finalise acquisition): Проверяется целостность и полнота полученных пакетов данных, фиксируются метаданные о процессе получения (время, объем).

Этап 5. Обработка (Process)

5.1. Интеграция данных: Данные СО интегрируются с внешними пространственными данными (границы административных районов, карты дорог) и другими источниками (административные регистры) для обогащения.

5.2. Классификация и кодирование: Пространственные события (ping с вышки) классифицируются и привязываются к географическим полигонам (город, район). Кодируются типы активности (ночная/дневная).

5.3. Редактирование и импутация: Применяются алгоритмические методы для «очистки» данных: фильтрация записей роуминга, коррекция "прыгающих" локаций, импутация пропусков с помощью методов машинного обучения.

5.5. Взвешивание и оценка: Проводится статистическое взвешивание для корректировки смещений выборки (например, недоучет детей и стариков, не имеющих телефонов). Создаются веса на основе калибровки по данным переписи населения.

Этап 6. Анализ (Analyse)

6.2. Валидация выходных данных: Валидация приобретает критически важный и сложный характер. Используются методы:

Перекрестная проверка: с данными переписи, регистров, опросов.

Анализ согласованности во времени: выявление аномальных всплесков или провалов.

Сценарный анализ: проверка устойчивости результатов к изменению параметров алгоритмов.

6.3. Интерпретация выходных данных и генерация пояснений: Требуются развернутые методологические пояснения о природе данных, ограничениях, потенциальных смещениях. Например, необходимо объяснить, почему «число активных SIM-карт» не равно «численности населения».

Этап 7. Распространение (Disseminate)

7.2. Производство продуктов распространения: Акцент смещается на машиночитаемые форматы (JSON, CSV через API) и интерактивные визуализации, позволяющие пользователям самим анализировать динамику.

7.3. Управление распространением: Реализуются механизмы дифференцированного доступа (например, открытые агрегированные данные и микроданные для исследователей по специальному соглашению).

Этап 8. Оценка (Evaluate)

8.1. Сбор отзывов пользователей: Пользователи - это часто технические специалисты и аналитики других ведомств. Сбор обратной связи фокусируется на удобстве API и качестве метаданных.

8.2. Оценка производственного процесса: Оценивается эффективность автоматизированного конвейера: скорость обработки, стоимость за 1 ТБ данных, частота сбоев.

Итоговый вывод: Использование данных СО трансформирует ТМПСИ из модели, ориентированной на дискретные, управляемые человеком процессы, в модель, описывающую непрерывный, высокоавтоматизированный, технологически насыщенный конвейер данных. Основные изменения происходят на этапах Проектирования (этап 2), где создаются сложные алгоритмы, Получения (этап 4), которое становится автоматическим, и Обработки/Анализа (этапы 5-6), где на первый план выходят методы машинного обучения и продвинутой статистической валидации для обеспечения качества и релевантности данных.

Внедрение Типовой модели производства статистической информации (ТМПСИ / GSBPM) в НСС стран СНГ

Общий тренд: Все национальные статистические службы (НСС) стран СНГ в той или иной форме ориентируются на ТМПСИ. Однако глубина, формальность и масштаб ее внедрения сильно различаются — от использования в качестве концептуального справочника до прямого внедрения в ИТ-системы и регламенты.

1. Казахстан:

Статус: ТМПСИ является официальной концептуальной основой для реинжиниринга статистических бизнес-процессов и развития ИТ-архитектуры.

Активная фаза внедрения началась в 2018-2020 годы в рамках реализации Концепции развития государственной статистики.

Модель и специфика:

Модель: ТМПСИ версии 5.х. Используется как каркас для описания "as-is" и проектирования "to-be" процессов.

Специфика:

1. Интеграция с ИТ: ТМПСИ легла в основу разработки и внедрения Единой платформы сбора и обработки данных (ЕПСОД) Бюро национальной статистики. Эта платформа фактически реализует этапы ТМПСИ в виде цифровых сервисов.

2. Акцент на этапах, связанных с «неопросными» данными: Особое внимание уделяется этапам «Проектирование получения данных» (Design data acquisition) и «Обработка» (Process) для интеграции больших данных.

2. Россия (Росстат):

Статус: Де-факто внедрена в ключевых процессах и ИТ-системах, хотя может не называться прямым аналогом ТМПСИ. Принципы модели глубоко интегрированы в методологию и новые технологические платформы.

Внедрение в явной форме с 2018 года.

Модель и специфика:

Модель: Фактически используется логика ТМПСИ, адаптированная к масштабам и структуре Росстата.

Специфика:

1. Отраслевая детализация: Для каждого направления статистики (сельхозперепись, экономическая статистика) разработаны детальные технологические схемы, которые являются отражением этапов ТМПСИ.

2. Централизация ИТ: Создание централизованных систем сбора и обработки (например, для экономической переписи) — это прямая реализация принципов ТМПСИ по стандартизации и сквозной автоматизации процессов.

3. Акцент на валидацию и контроль: Этапы «Обработка» и «Анализ» (Process, Analyse) в российской практике чрезвычайно детализированы и формализованы, что соответствует национальной специфике требований к качеству.

3. Беларусь (Белстат):

Статус: Используется как методологический ориентир для совершенствования процессов и гармонизации с международными стандартами.

Внедрение с 2015 года.

Специфика:

Тесная связь с системой качества: Внедрение ТМПСИ тесно увязано с внедрением СТБ 8.0 и внутренних инструкций по качеству. Каждый этап ТМПСИ сопровождается контрольными точками по качеству.

4. Узбекистан и Кыргызстан:

Статус: Активное знакомство и пилотное внедрение при поддержке международных доноров (Всемирный банк, ПРООН, Евростат).

Внедрение в рамках проектов, стартовавших после 2020 года.

Специфика:

1. Обучение и наращивание потенциала: Основной фокус — на обучении сотрудников НСС принципам и терминологии ТМПСИ.

2. Пилотные применения: Модель применяется для редизайна конкретных процессов, например, для организации сбора данных по коротким опросам предприятий или для описания процесса использования данных ДЗЗ.

5. Вызовы:

Недостаток ИТ-инфраструктуры: Полная реализация ТМПСИ требует современных гибких ИТ-систем.

Кадровый вопрос: Нехватка специалистов, которые понимают сквозной статистический бизнес-процесс.

Вывод: ТМПСИ в странах СНГ стала практическим инструментом модернизации, который, с разной степенью интенсивности, используется для повышения эффективности, стандартизации и качества статистического производства.