Региональный учебный семинар по переходу к методам проведения переписей населения и жилищного фонда на основе регистров

Анкара, Турция

12-15 июня 2023 г.

Интеграция данных и преобразование административных данных в данные переписи

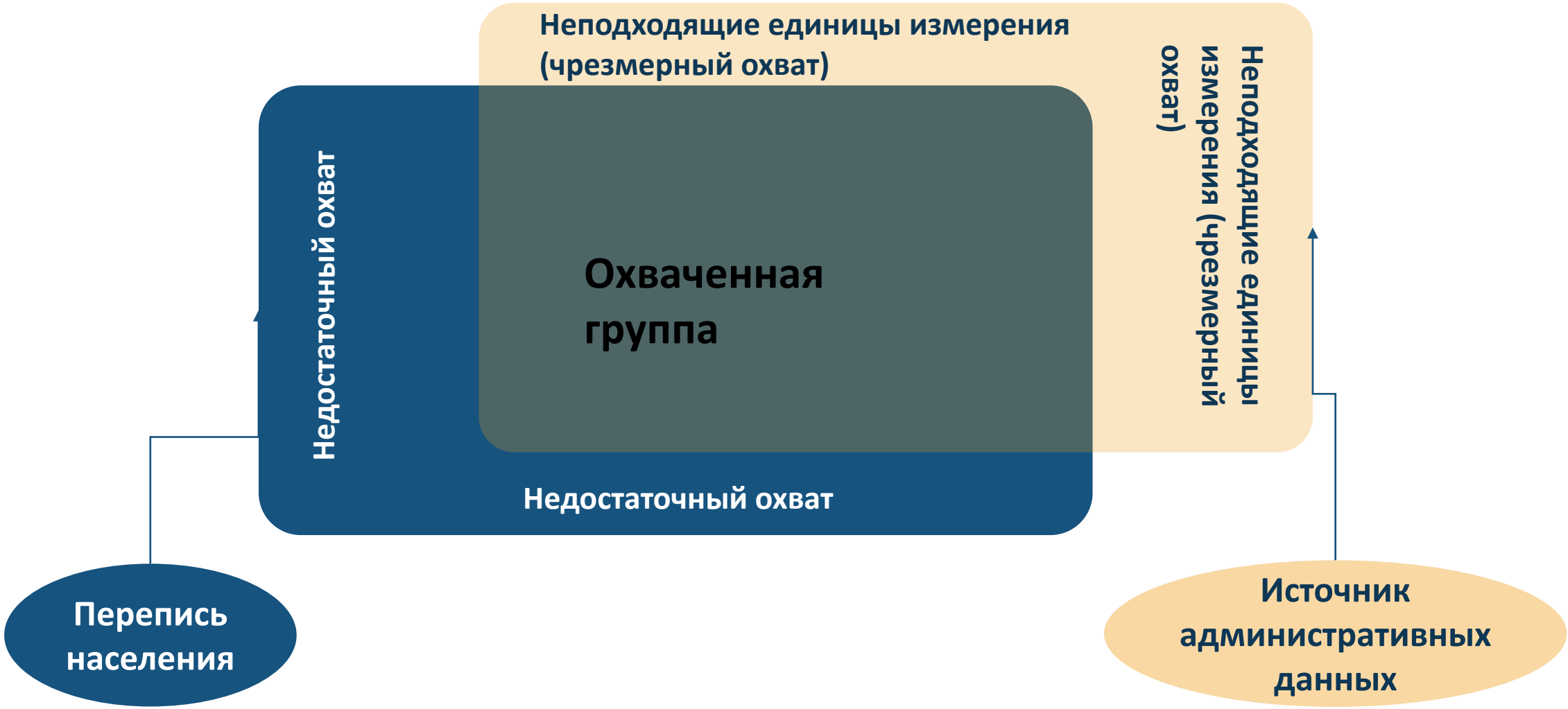
Региональный учебный семинар по переходу к методам проведения переписей населения и жилищного фонда на основе регистров

12-15 June 2023, Ankara, Türkiye

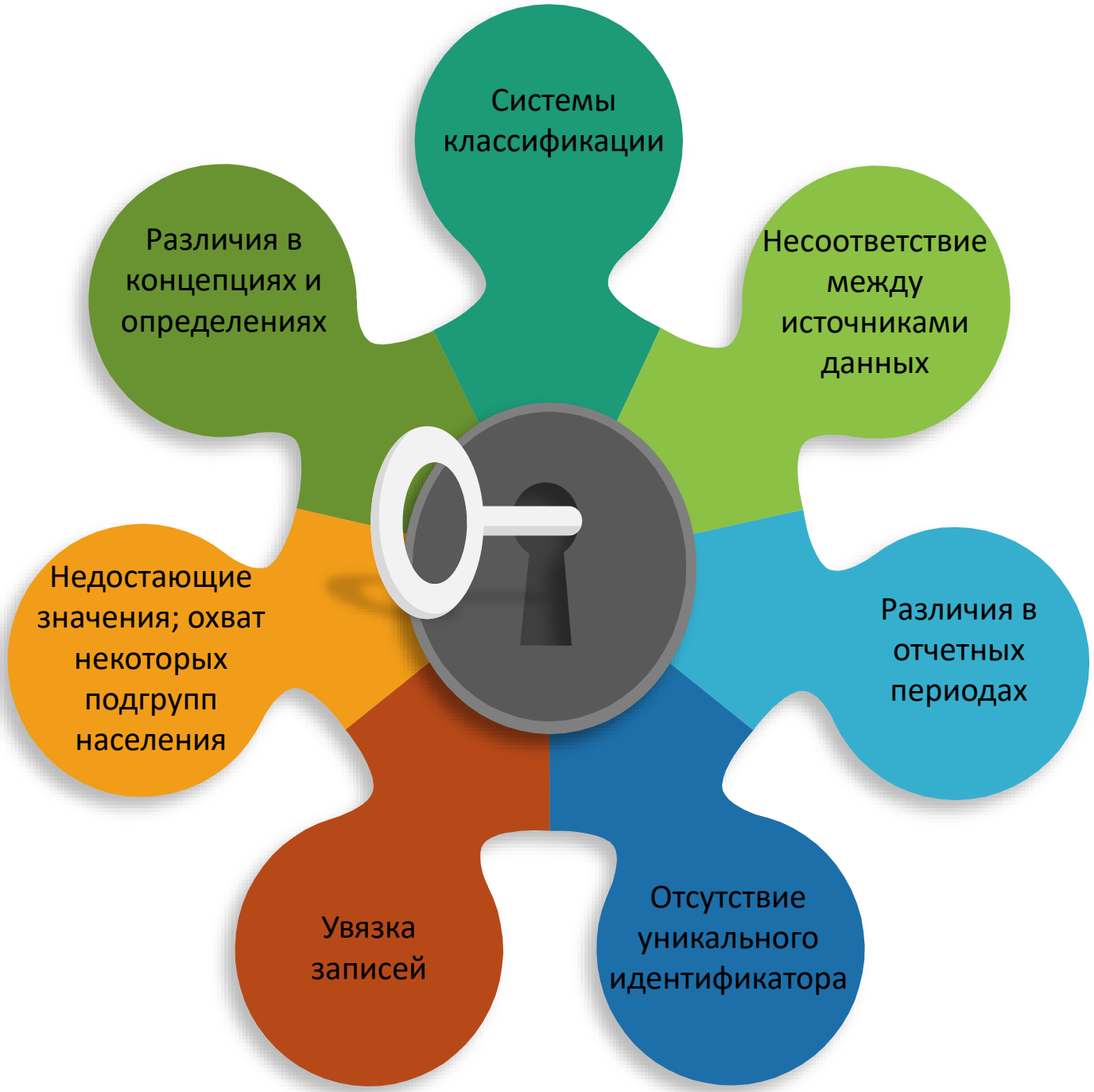
Афсане Яздани, Статистический отдел ЭСКАТО

Общая модель перехода от традиционной переписи к методам, основанным на регистрах

Охват населения переписью



Некоторые из ключевых технических трудностей и вопросов, связанных с использованием административных источников данных



Создание интегрированных статистических регистров для целей переписи

- В дополнение к определению источников данных, которые будут использоваться, некоторыми из ключевых процессов, связанных с составлением статистического регистра, являются:
  - Увязка данных
  - Устранение дублирования данных
  - Устранение случаев противоречивой информации
  - Обновление и метод «Признаков жизни»
  - Редактирование и вменение

Увязка данных

- Увязка записей относится к идентификации и объединению записей, соответствующих одним и тем же объектам – например, лицам, предприятиям, жилищам и домохозяйствам – в двух или более источниках данных.
  - **Детерминированное или точное сопоставление** - это когда в источниках данных существует формальное правило принятия решения, обычно в виде уникальных идентификаторов, таких как личные идентификационные номера (PIN-код), которые будут использоваться для сопоставления.
  - **Вероятностное сопоставление** - это когда строгие правила принятия решений неприменимы. Вместо этого устанавливаются сложные вероятностные правила принятия решений на основе набора ключевых переменных, которые являются общими в источниках данных, таких как имя, пол, дата рождения и адрес, для присвоения оценок сходства.
  - Может быть применен **комбинированный метод,** при котором сначала используется точная/детерминированная привязка для как можно большего числа записей, а затем вероятностная привязка для остальных записей.

1. Увязка данных – идентификационные номера

- **«Уникальные идентификационные номера2** значительно облегчают увязку нескольких источников административных данных.
  - Уникальный идентификатор должен быть общим для всех соответствующих регистров.
  - Идентификационные номера часто создаются в административных целях для использования в регистрах населения, записях актов гражданского состояния, национальных системах идентификации или других административных регистрах. Хотя иногда они вводятся в статистических целях.
  - Способы внедрения идентификационных номеров различаются в разных странах. Иногда номер относится к атрибутам личности, а иногда является уникальным номером, не содержащим информации. Иногда включают контрольные цифры. Он может быть выдан гражданам по достижении ими совершеннолетия или при рождении.
  - Для повышения защиты персональных данных идентификационные номера зашифрованы, чтобы предотвратить считывание информации неавторизованными лицами.

1. Увязка данных – установление связи между лицами и жилищами/домохозяйствами

- Основными расчетными единицами переписи являются лица, домохозяйства, семьи и жилища; это следует учитывать при переписи, основанной на регистрах:
  - Минимально необходимыми идентификаторами являются данные о лицах (PIN-код) и местах проживания (код адреса и/или пространственные координаты). Каждому проживающему в нем лицу присваивается идентификационный номер жилого помещения.
  - При традиционной переписи домохозяйства формируются с использованием “концепции ведения домашнего хозяйства”. Это является сложной задачей при переписи на основе регистров; таким образом, многие страны вместо этого используют “концепцию домашнего жилья”, которая рассматривает всех лиц, проживающих в одной и той же жилищной единице, как домашнее хозяйство.
  - В некоторых странах существует отдельный регистр домашних хозяйств, который облегчает процесс создания домашних хозяйств.

2. Устранение дублирования данных

- Чтобы избежать серьезных проблем с охватом, должен быть внедрен адекватный процесс обнаружения и устранения дублирования. Дублирование лиц в статистическом регистре населения может иметь место, если:
  - Надлежащие методы увязки данных отсутствуют.
  - Идентификаторы в источниках данных имеют низкое качество, что приводит к ложному совпадению/несоответствию записей.
  - Повторно созданные единицы (действительно родившиеся или иммигранты) и удаленные единицы (действительно умершие или эмигранты) плохо отражаются в источниках административных данных.

Вести журнал изменений регистров полезно.

3. Устранение случаев противоречивой информации

Потенциальные причины:

- задержка с представлением информации об изменениях, внесенных отдельными лицами,
- задержка в обновлении административными органами,
- инциденты с несколькими домами,
- различные определения, классификации или ошибки в отчетных периодах
- ошибки в одном источнике.

Для разрешения проблем

- Определите, какой источник с наибольшей вероятностью будет обновлен и является точным для конкретной переменной
- Установите приоритет/правило принятия решения для каждой переменной
- Убедитесь, что источники с более низким приоритетом не перезаписывают данные из источника с более высоким приоритетом.

Негативные последствия

- Неадекватные правила приоритета/принятия решений приводят к ошибкам содержимого
- Противоречивая (или множественная) адресная информация может привести к ошибкам недостаточного или чрезмерного охвата, особенно на субнациональном уровне.

4. Обновление и метод «Признаков жизни»

- Основным минимальным источником информации для поддержания SPR в актуальном состоянии является информация, полученная в результате регистрации актов гражданского состояния о рождениях, смертях, браках и любых изменениях адреса в результате внутренней или международной миграции.

SPR должен охватывать население, охваченное переписью, т.е. включать только тех лиц, которые живы и отвечают набору заранее определенных критериев проживания.

Признаки жизни (SOL) - это широко используемый инструмент, помогающий минимизировать почти неизбежные недостатки охвата в SPR.

4. Обновление и метод «Признаков жизни»

- Признаки жизни (SOL) - это набор «правил активности», которые можно использовать для проверки различных административных источников данных, доступных НСУ, чтобы определить, жив ли человек и проживает ли он в определенный период времени
- Список маркеров SOL никогда не может быть абсолютным; однако, чем больше маркеров можно использовать, тем точнее будет оценка.
- Маркеры SOL могут быть определены с использованием данных из РАГС, налоговых регистров, социального страхования, базы данных по безработице, базы данных об образовании и т.д.

Если человек хотя бы раз был активен (имеет учетную запись) в регистре в течение определенного года, то значение SOL для него/нее равно 1; в противном случае - 0.

4. Обновление и метод «Признаков жизни»

- Чтобы использовать эти маркеры, необходимо выполнить два ключевых предварительных условия, по крайней мере на национальном уровне:

  a) следует вести ряд административных регистров, в которых

  - все лица во всех регистрах идентифицируются по их уникальным идентификационным кодам; и
  - все жилые помещения (жилища, семейные дома и т.д.) во всех регистрах идентифицируются по их уникальным идентификаторам адресов;

  b) все регистры должны охватывать все население и регулярно обновляться, по крайней мере, ежегодно

5. Редактирование и вменение

- В идеале SPR должен быть чистым и согласованным, без противоречий между отдельными элементами данных, без пропущенных или неправдоподобных значений.
  - Административные источники данных должны быть тщательно исследованы для выявления и устранения «систематических ошибок» (либо ошибок в охвате, либо ошибок в содержании).
  - Наличие метаданных, особенно о процедурах редактирования в рамках административного органа, очень полезно, особенно для понимания существования каких-либо систематических ограничений.

Методы решения проблем

- Редактирование: для исправления значений, которые явно ошибочны или неправдоподобны
- Условное вычисление: для вставки правдоподобных значений там, где отсутствуют элементы данных
- Одновременное редактирование и проверка согласованности нескольких источников данных
- Тесный контакт с административными органами для улучшения способа сбора и записи данных.

Исследования и тестирование

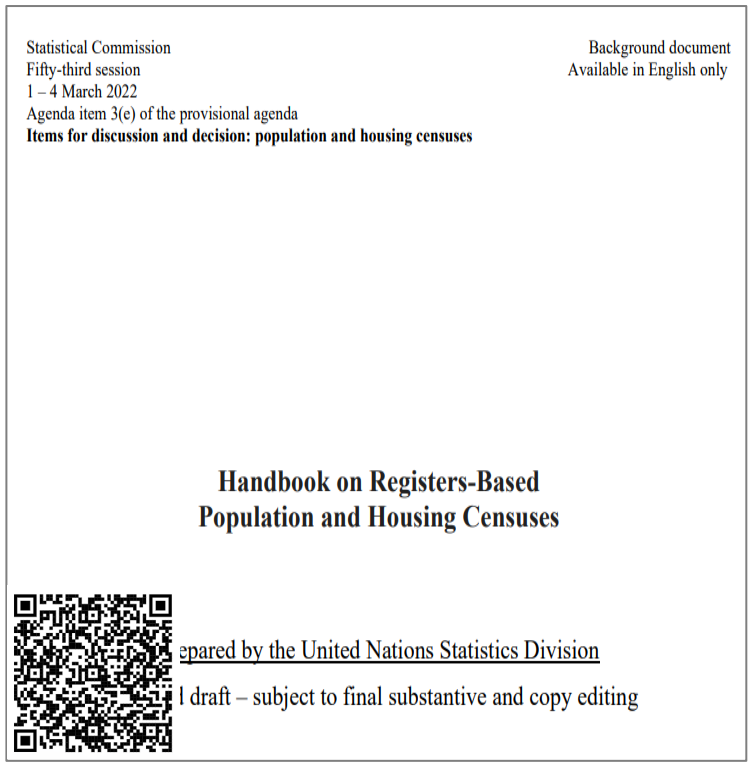
- Никогда не включайте новые административные источники данных в процесс подготовки данных переписи без «технико-экономического обоснования» со стороны НСУ. Технико-экономическое обоснование включает:
  - формирование детального представления о процессах сбора данных административным органом, охватываемом населении и переменных, включенных в административный источник, а также о том, насколько доступны эти данные.
  - детальное изучение данных испытаний для выявления проблемы с качеством и определения очистки и гармонизации, а также проверки валидации.
  - объединение с другими доступными регистрами для проверки качества данных и выбора наиболее надежных переменных и значений в соответствии с разработанными методологическими правилами.
  - подготовку оценок с использованием тестовых данных и их оценку путем сравнения с результатами предыдущих переписей или другими источниками.
  - использование технико-экономического обоснования для разработки методов получения характеристик переписи

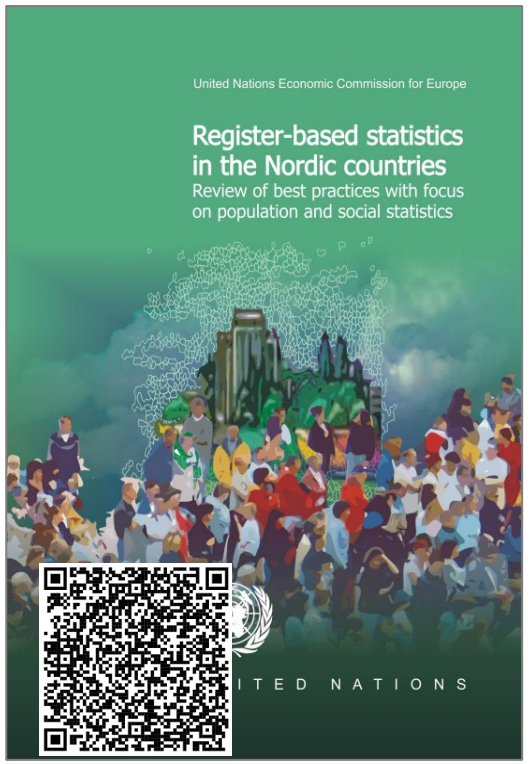
Исследования и тестирование

- НСУ должны решать следующие основные задачи при получении характеристик переписи.
  - определить международный стандарт переписи (определение, классификация и т.д.), применимый к целевой характеристике переписи;
  - сравнить определения и классификации переписи с определениями и классификациями, используемыми в административном источнике;
  - проверять точность административных данных, записанных из альтернативных источников, и сотрудничать с поставщиками данных для устранения/смягчения любых недостатков;
  - определить, какие источники и в каком количестве требуются для получения и обеспечения качества каждой целевой характеристики переписи;
  - установить оптимальные правила для получения каждой характеристики переписи и разработать необходимое программное обеспечение для обработки данных, оптимизированное с учетом качества запрашиваемых результатов;
  - а в тех случаях, когда характеристики не охватываются никакими административными источниками, предпринять шаги для обеспечения создания необходимого регистра или его части (например, предложить поправки к процедурам регистрации, правовой среде и т.д.).

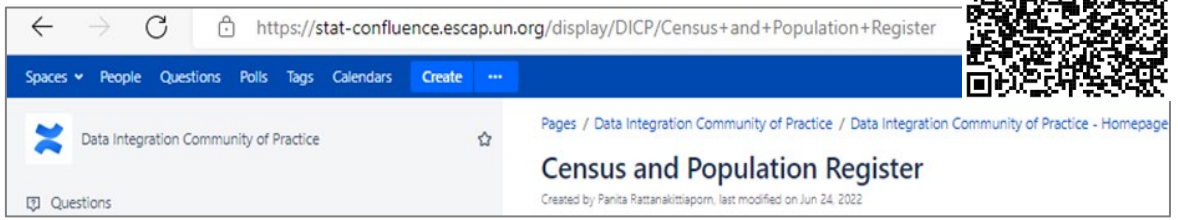
Выводы

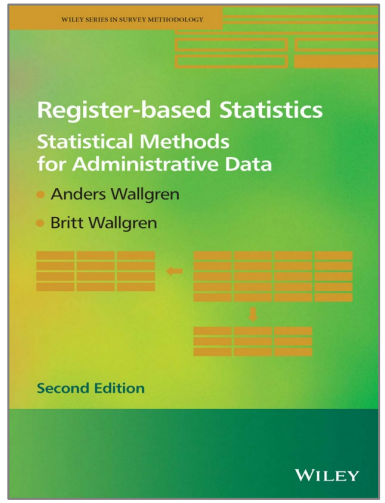
- Создание статистического регистра населения (SPR) - сложный процесс, включающий в себя несколько этапов, которым необходимо следовать с особой осторожностью.





Сообщество специалистов по интеграции данных (DI-CoP)






Thank
you