United Nations | Department of Economic and Social Affairs

Оценка качества источников административных данных

Мерьем Демирчи
Статистический отдел ООН

Оценка качества

- **Неотъемлемая часть проведения переписи,** независимо от типов методологий переписи
- **Всеобъемлющий процесс,** охватывающий все этапы переписей - качество одного этапа оказывает влияние на качество следующего этапа
  - ✓ Процесс оценки качества разработан по-другому для переписей, которые проводятся с использованием административных источников данных, по сравнению с традиционной переписью

Что мы узнаем?

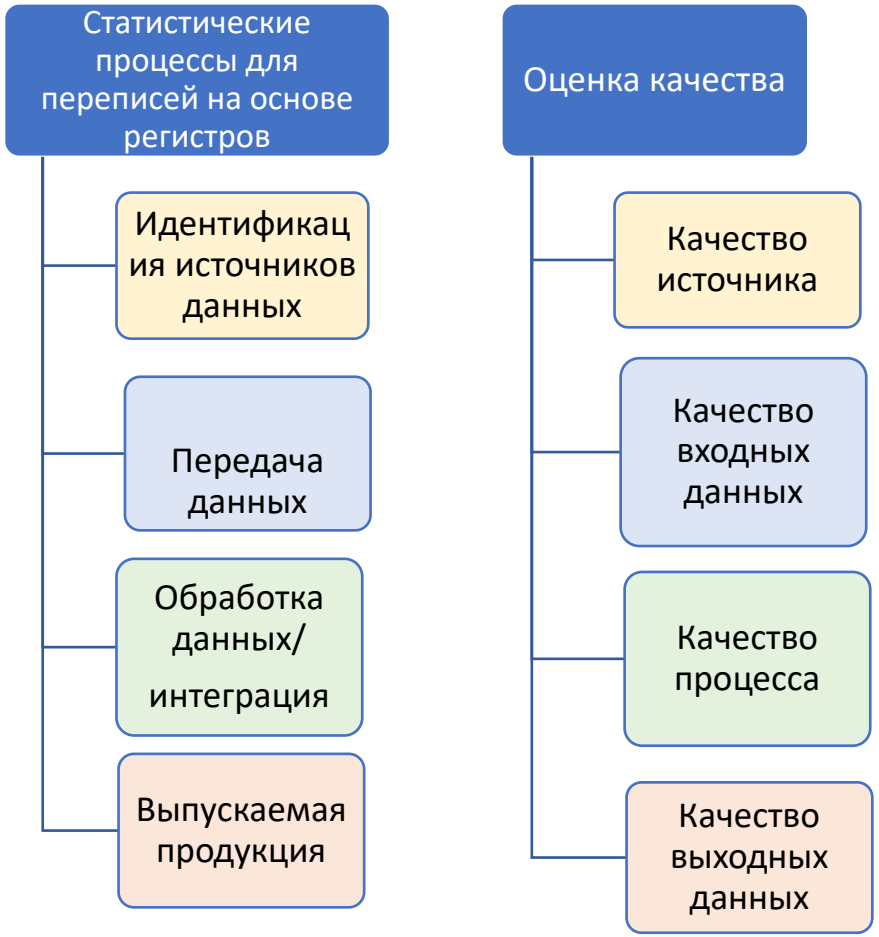
- ❑ **Как спроектировать процесс оценки качества?**
- ✓ **Этапы** оценки качества административных источников данных
- ❑ **Что проверять на каждом этапе?**
- ✓ **Измерения** (компоненты) оценки качества
- ❑ **Как измерить качество?**
- ✓ **Показатели** для измерения качества

Этапы оценки качества

Четыре этапа

Этапы оценки качества административных источников данных

- Этапы оценки качества административных регистров и полученных на их основе переписных данных в целом соответствуют этапам статистических процессов переписи
- Разработка процесса оценки качества на основе этих четырех этапов поможет обеспечить, чтобы оценки переписи основывались на наиболее подходящих источниках и методах



Аспекты оценки качества

Релевантность-степень, в которой статистические результаты отвечают текущим и потенциальным потребностям пользователей с точки зрения доступности данных, концепций и определений

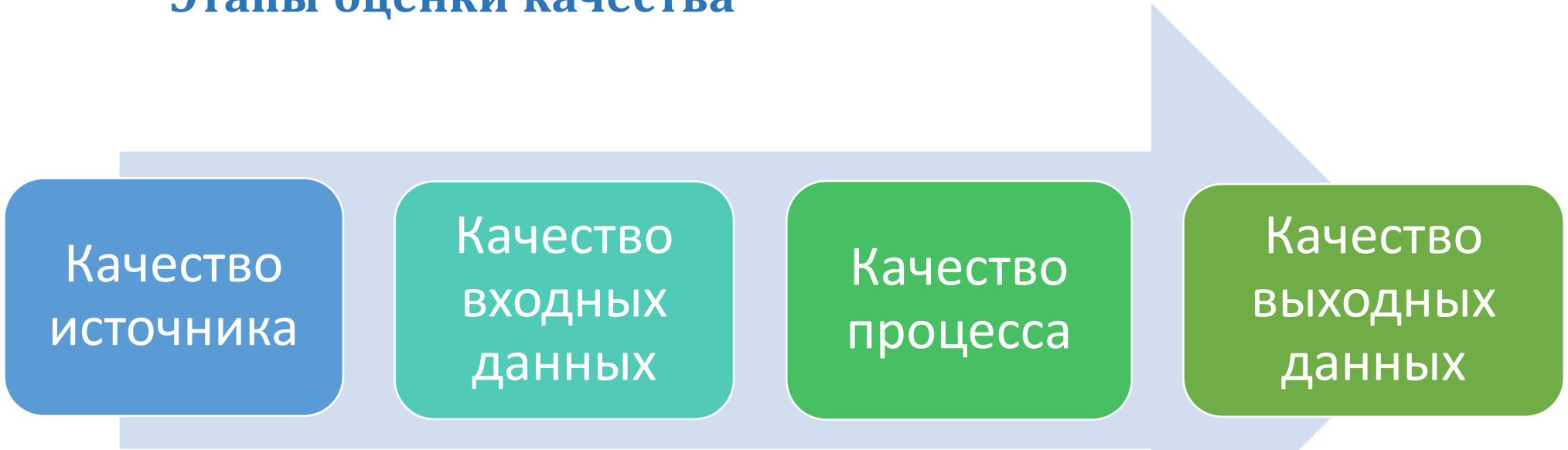
Точность и достоверность - степень, в которой информация правильно описывает явления, такие как обычное постоянное население

Актуальность - задержка между датой, к которой относятся данные (день переписи), и датой, когда информация становится доступной

Согласованность и сопоставимость - степень сходства данных, полученных из различных источников или методов, – степень сопоставимости данных с течением времени - степень согласованности между источниками данных и временем

Доступность и интерпретируемость - легкость, с которой пользователи могут получить доступ к данным переписи – наличие метаданных, описывающих источники, методы и определения

Этапы оценки качества



-Оценка качества источников **административных данных** *Оценка на основе метаданных*

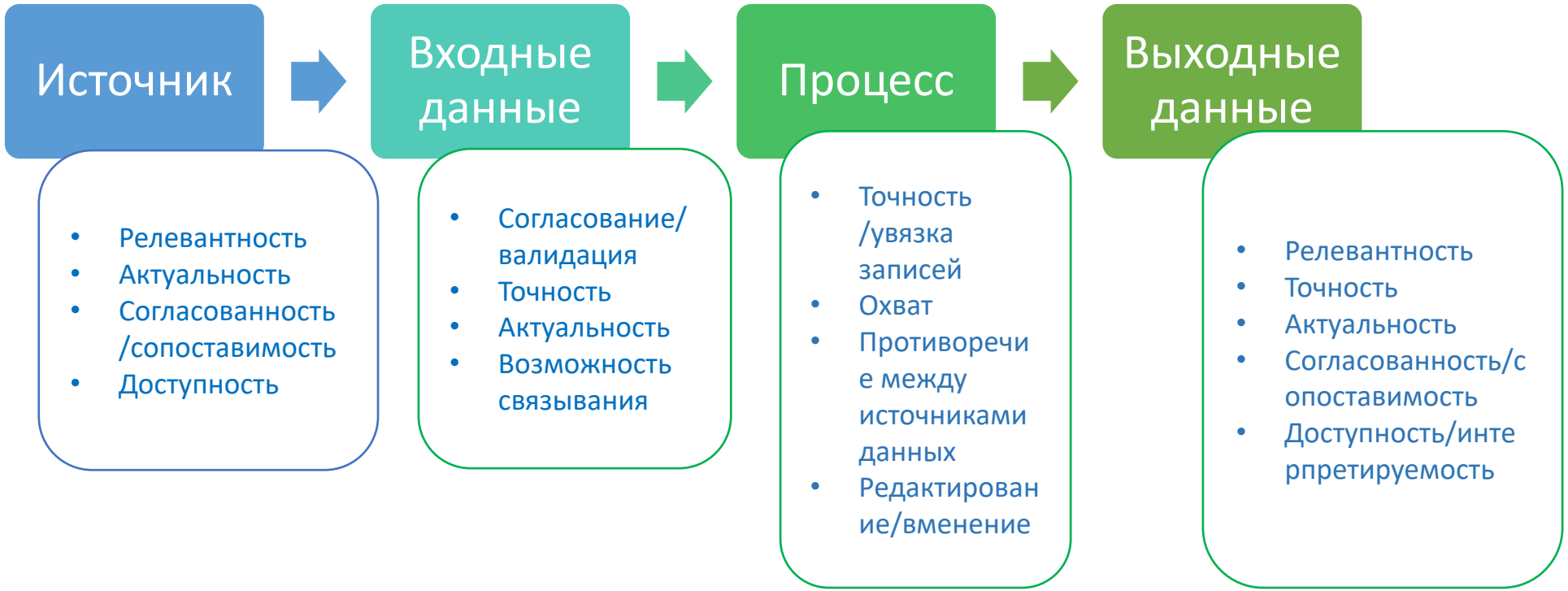
-Оценка качества **необработанных административных данных**, *поскольку они предоставляются НСУ административными органами*

- Оценка изменений в качестве данных, возникающих в **результате интеграции данных и обработки административных данных**

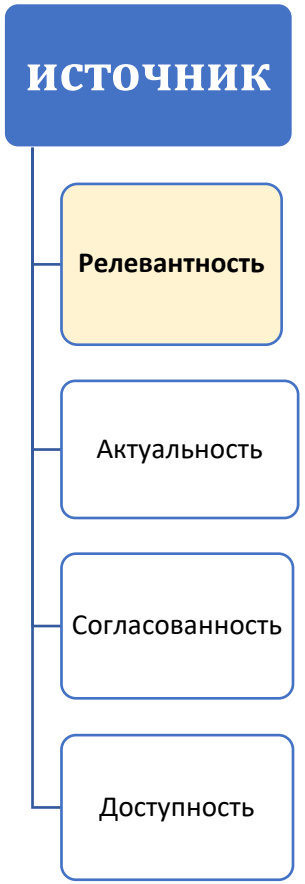
- Общая оценка качества статистических результатов, распространяемых среди пользователей

United Nations | Department of Economic and Social Affairs

Параметры качества для каждого этапа



Качество источника (1)



Оценка **ошибок представления и измерений**

- **Ошибка представления**
  - Приведение единиц в регистре в соответствие с целевыми единицами переписи (лица, домохозяйства, жилые единицы)
  - Информация о том, какие **законы и нормативные акты** определяют, кто будет включен/исключен из административных источников данных
  - Информация о том, какие **методы/процедуры** используются для включения/обновления/исключения единиц измерения

Индикатор оценки

* Соответствует ли охват регистра населения потребностям переписи?

* данные о недостаточном и/или избыточном охвате – провести оценку всех групп населения, которые должны быть включены в регистр населения

Качество источника

- **Ошибка измерения**
  - Согласование концепций и определений переменных в регистрах с концепциями и определениями признаков переписи

* включает ли регистр переменные, необходимые для переписи, или нет?

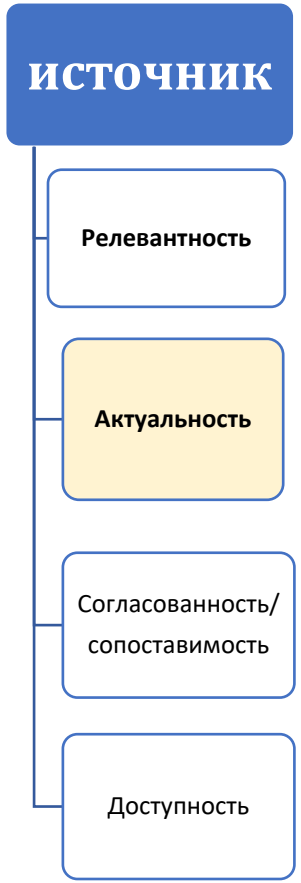
*соответствуют ли административные концепции, определения и классификации для таких переменных тем, которые были приняты в ходе переписи?

*в случае несоответствия, возможно ли преобразование переменных для удовлетворения требований переписи?

*если это невозможно, предоставит ли это аналогичную информацию?

United Nations | Department of Economic and Social Affairs

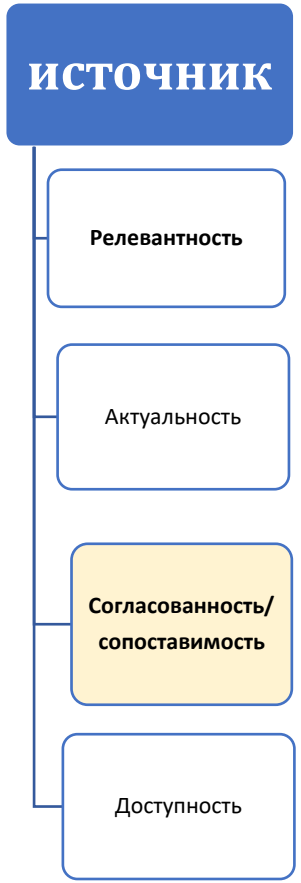
Качество источника-Актуальность



- Разница между исходной датой, к которой относятся данные, и датой их предоставления в НСУ - чем больше задержка, тем менее актуальность
- Некоторые примеры информации, которая может быть использована для оценки своевременности

*Каков временной промежуток между датой возникновения и датой регистрации?
*Каков временной интервал между датой регистрации и датой предоставления данных в НСУ?
*Был ли регистр полностью обновлен при его предоставлении НСУ или нет?
* Как часто данные могут предоставляться в НСУ для обновления или получения информации о новых лицах или жилищах?

Качество источников - согласованность и сопоставимость



- оценить степень, в которой административный источник может быть **успешно связан и объединен с другими источниками** данных для использования в переписи

- Некоторые примеры информации, которая может быть использована для этой оценки :

  * Содержит ли источник уникальный идентификатор (например, PIN-код), который является общим с уникальным ключом, необходимым для привязки переписи?
  * Если да, то доступен ли идентификатор для всех соответствующих лиц/адресов в источнике или только для особых групп населения или географических районов?
  * Содержит ли источник уникальную комбинацию переменных (таких как имя, дата рождения и адрес), которые можно было бы использовать для увязки переписи?

Качество источника - доступность и интерпретируемость

- Важно **определить любые ограничения, которые могут повлиять на способность НСУ приобретать и использовать административный источник**, такие как существующие ограничения по защите данных

* *Каков уровень общественной приемлемости?*

Решение о том, получит ли НСУ доступ к определенному источнику данных для использования в переписи или нет, также может зависеть от общественного признания

* *Насколько легко передавать данные?*

Поставщик данных может использовать модели данных, форматы, схемы, программное и аппаратное обеспечение, сильно отличающиеся от тех, с которыми знакомо НСУ

* *Существуют ли четкие и всеобъемлющие метаданные?*

Оценка интерпретируемости связана с наличием всеобъемлющих и понятных метаданных и документации об административном источнике

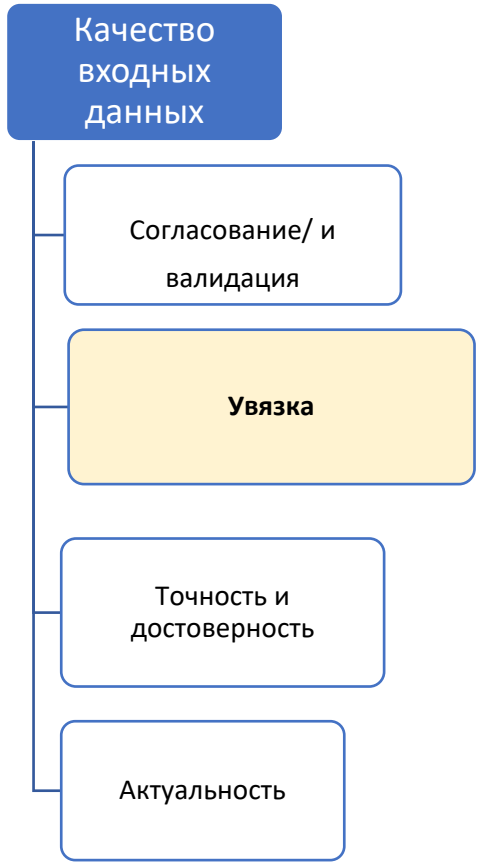
Качество входных данных - согласование и валидация

- Для НСУ крайне важно обеспечить, чтобы передаваемые файлы данных были в требуемом "читаемом" формате; базы данных структурированы таким образом, чтобы системы НСУ могли принимать и считывать их

Для оценки действительности могут быть использованы некоторые показатели, в том числе:

* Правильно ли названы и отформатированы указанные переменные (например, числовая, категориальная, текстовая информация и т.д.).,
* Был ли указан правильный отчетный период или нет
* Соответствуют ли переменные ожидаемому предопределенному содержимому или нет, устанавливается с помощью метаданных, собранных на этапе создания источника

Качество входных данных - Возможность увязки

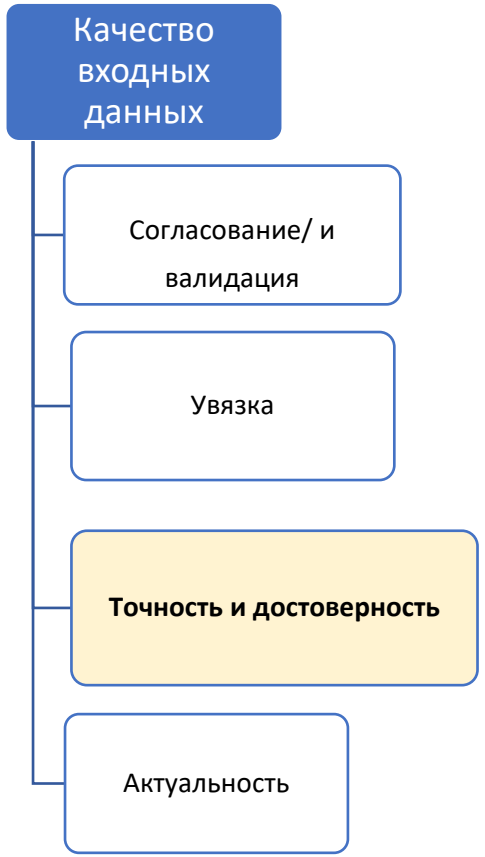


- Оценка переменных в каждом источнике административных данных, используемых при увязке, – информирование о разработке успешной увязки на этапе процесса

Для оценки увязки могут быть использованы некоторые показатели, в том числе:

* Процент уникальных значений или комбинация переменных, которые будут использоваться при увязке
- например, процентное соотношение уникального персонального идентификационного номера или комбинации возраста, даты рождения и адреса
* Ошибки измерения в пределах переменных связи
-Процент пропущенных значений, неправдоподобных значений и т.д.
* Распространенность предвзятого распределения
- существует ли значительно более высокая доля значений вне диапазона или отсутствующих значений для ключевой переменной (переменных) связи в определенных географических регионах

Качество входных данных - точность и достоверность

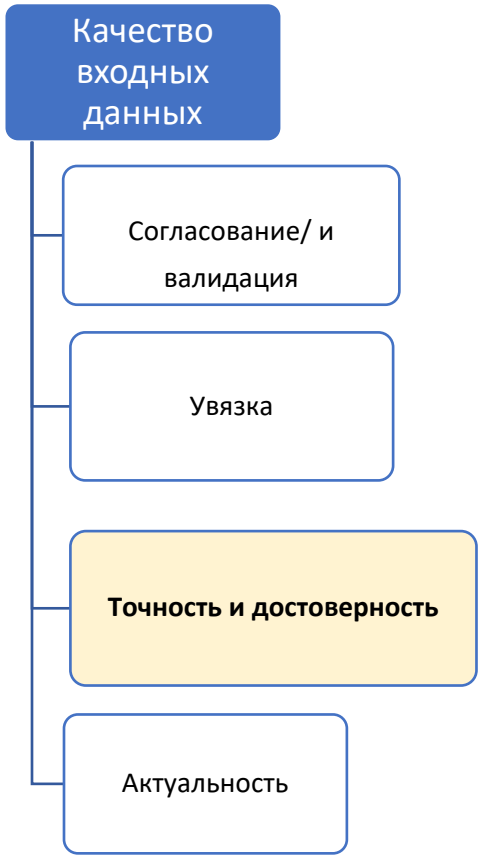


- При оценке точности входных данных НСУ следует проводить различие между
  - **"репрезентативными ошибки"** (относящиеся к охвату целевой группы населения) и
  - **"ошибками измерения"** (относящиеся к конкретной рассматриваемой переменной).

- Основные показатели для оценки **репрезентативных ошибок** включают:
  - общее количество полученных единиц (человек/единиц жилья) (для сравнения с ожидаемым количеством);процент повторяющихся единиц измерения
- Ключевым показателем при оценке **недостаточного охвата** было бы:
  - процент единиц в справочном источнике (традиционная перепись или полный базовый регистр), которые отсутствуют в предоставленном (административном) источнике.

  в то время как **чрезмерный охват** может быть оценен по:
  - проценту единиц в (предоставленном) источнике, не принадлежащих к целевому постоянному населению НСУ

Качество входных данных - точность и достоверность



- Оценка **ошибок измерений**
- Основные показатели для измерения полноты характерных переменных, предоставляемых в наборах административных данных на агрегированном уровне (таких как возраст, пол, этническая принадлежность и т.д.), включают следующее:

  - количество и процент пропущенных значений в ключевых переменных (таких как дата рождения и пол);
  - количество и процент значений, выходящих за пределы диапазона, в пределах ключевых переменных (например, зарегистрированный возраст 120 лет);
  - количество и процент неправдоподобных значений (на основе, например, перекрестных таблиц различных переменных);
  - распространенность неожиданных частот, закономерностей или выбросов, основанная на анализе частоты/распределения ключевых переменных

Качество вводимых данных-актуальность

Показатели **актуальности** можно относительно легко определить путем сравнения

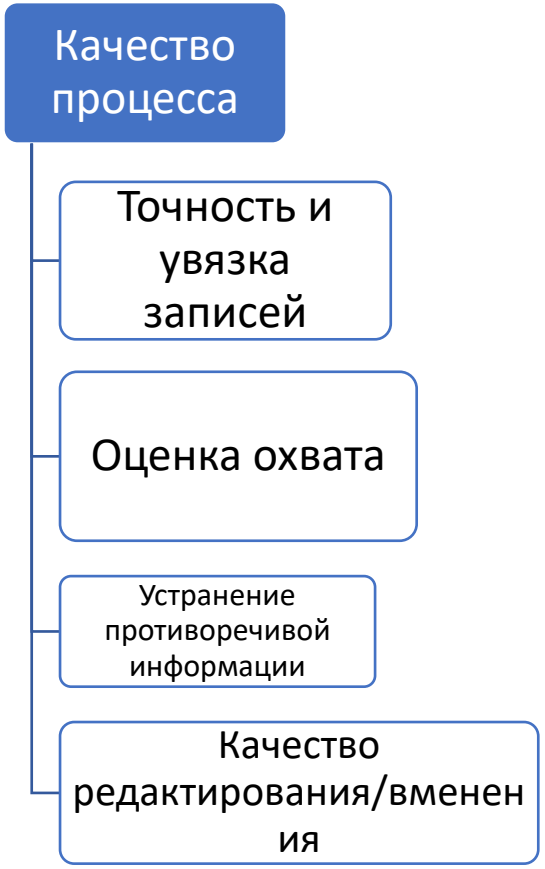
- исходная дата,
- указанная дата поставки, и
- фактическая дата доставки данных

Два аспекта своевременности можно оценить следующим образом:

- **разница между датой возникновения и регистрацией**
  - дата регистрации владельцем регистра любых изменений в источнике данных и дата фактического изменения в совокупности;
- **разница между датой получения данных** НСУ и датой базисного периода, к которому относятся данные,

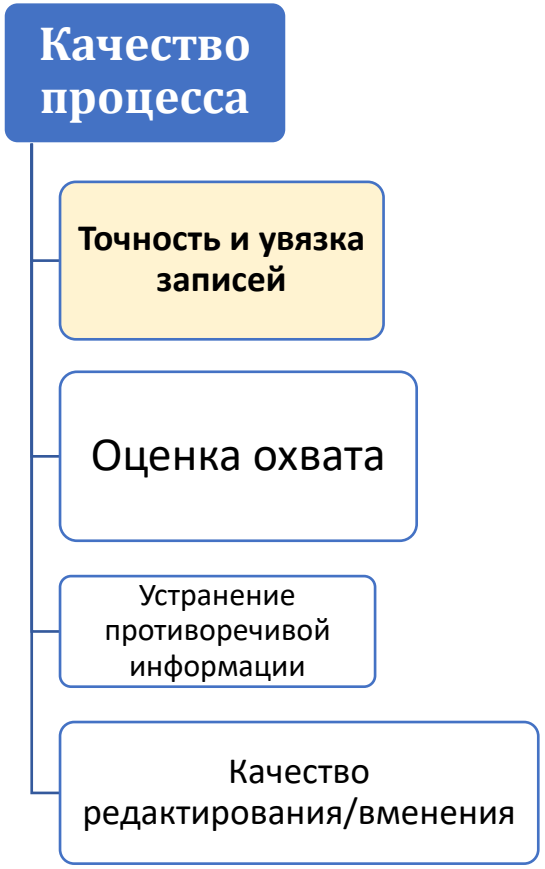
чем дольше задержка, тем меньше актуальность

Качество процесса



- Поскольку данные, хранящиеся в административном источнике, собираются не для статистических целей, НСУ должны каким-либо образом преобразовать их для использования в переписи.
  - Увязка данных с помощью общего идентификатора
  - Создание/обновление статистического регистра населения
  - Обработка данных
    - Устранение дублированной даты
    - Устранение противоречивой информации
    - Метод обновления и выявления признаков жизни для улучшения качества охвата статистического регистра населения
    - Редактирование и вменение
    - Проверка достоверности результатов переписи

Качество процесса -точность и увязка записей



- Распространенными методами оценки качества увязки являются:
  - определение доли записей, которые не связаны или не могут быть связаны между собой,
  - количество и процент повторяющихся ключей увязки
  - отсутствие ключей увязки или количество пропущенных или неправдоподобных значений
    - Сравнение распределений характеристик связанных и несвязанных записей, например, по таким переменным, как возраст и пол, а также по регионам и подгруппам населения
      - Различия в характеристиках предполагают, что некоторая неточность возникает из-за ошибки увязки,

Качество процесса - Оценка качества статистического регистра населения

- *Использование признаков жизни*
  - Все чаще используемым инструментом, помогающим свести к минимуму чрезмерный охват, является так называемый метод "признаков жизни" (SOL), основанный на ряде "правил", которые приняты для обеспечения того, чтобы в перепись включались только те лица, которые живы и отвечают набору заранее определенных критериев проживания
- *Использование независимых опросов и демографического анализа*
  - В дополнение к подходу, основанному на признаках жизни, существует несколько других методов, доступных для оценки охвата переписей (и, действительно, ошибки в содержании).
  - К ним относятся: простые методы обеспечения качества, такие как внутренние проверки согласованности; демографический анализ; сравнения с данными из других источников, включая предыдущие переписи и/или текущие обследования домашних хозяйств

Качество процесса

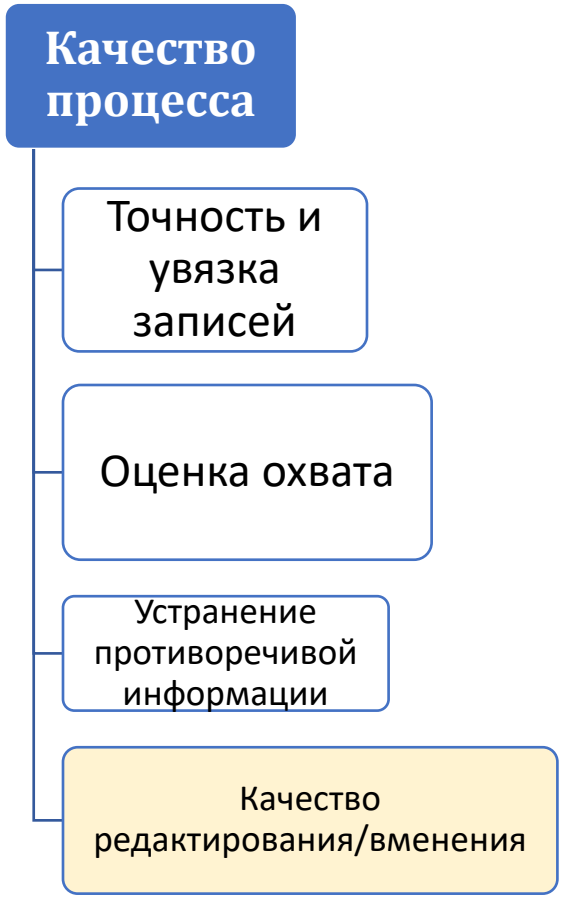
- Качество процесса
  - Точность и увязка записей
  - Оценка охвата
  - Устранение противоречивой информации
  - Качество редактирования/вменения

Устранение противоречивой информации

- При использовании данных из нескольких источников требуются методы оценки качества переменных, когда одна и та же переменная указывается в разных регистрах
  - Правила необходимы для определения того, какое значение является точным

- Разрешить проблему с противоречивой информацией можно по адресу проживания, зарегистрированному в разных регистрах, и определите наиболее точную информацию
  - Затем НСУ необходимо будет решить, к какому адресу должна относиться информация переписи (возможно, используя, например, самый последний указанный адрес).

Качество процесса



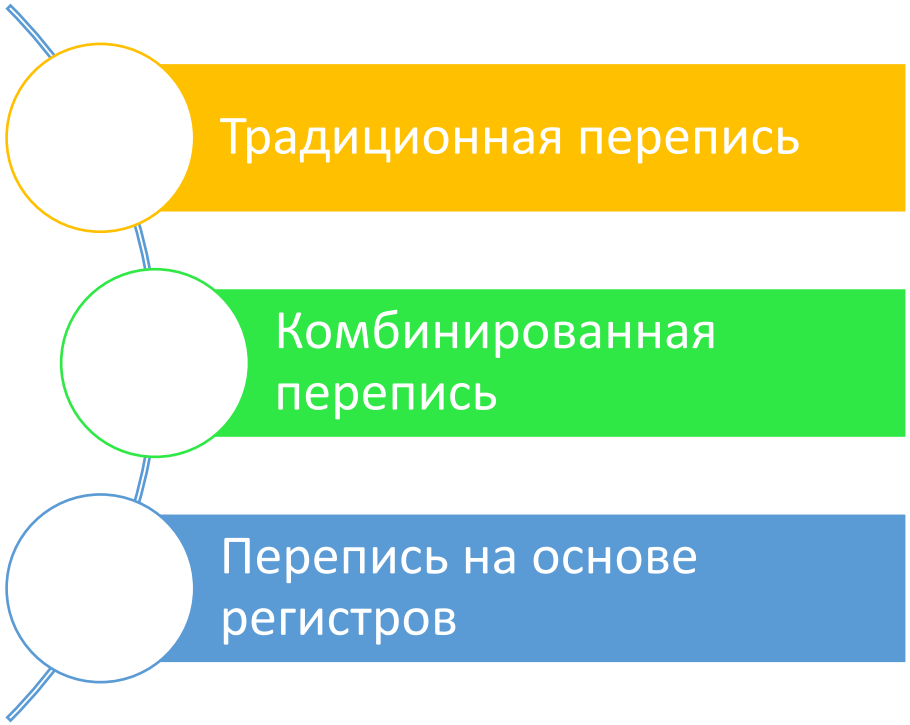
Оценка качества процесса редактирования и вменения

- Редактирование и условное вычисление - это итеративный процесс просмотра данных с целью исправления любых ошибок, возникающих в результате недопустимых, несогласованных или неправдоподобных значений, а для недостающих значений
  - может использоваться ряд показателей, таких как частота сбоев редактирования, частота корректировок, частота условных вычислений, индекс несходства

Качество выходных данных

- Независимо от типа методологии сбора данных, используемой для переписи, для НСУ одинаково важно оценить качество результатов

- Для тех НСУ, которые перешли на основанную на регистрах или комбинированную методологию сбора данных переписи, особенно важно оценить качество результатов, определить, повлиял ли переход на общее качество результатов



Выводы

- Каждая страна должна планировать процесс перехода на основе
  - доступности источников административных данных
  - оценки качества источника административных данных и качества входных данных

- Переход следует планировать постепенно,
  - с каждым разом вводя все больше источников административных данных и переменных, при условии, что качество регистров будет подтверждено

- В результате перехода могут произойти некоторые изменения в определениях переменных, совокупных базах данных и классификациях результатов
  - Следует оценить влияние этих изменений на качество статистических данных и разъяснить результаты пользователям

United Nations | Department of Economic and Social Affairs

Справочные документы

Руководящие принципы ЕЭК ООН по оценке качества административных источников для использования в переписях https://unece.org/statistics/publications/CensusAdminQuality

Справочник СООН по переписям населения и жилищного фонда на основе регистров https://unstats.un.org/UNSDWebsite/statcom/session_53/documents/BG-3e-Handbook-E.pdf

Спасибо за внимание…

Email: demircim@un.org