Food and Agriculture Organization of the United Nations

Вебинар по практическим аспектам Программы всемирной сельскохозяйственной переписи 2020 года (ВСП-2020)

Виртуальная встреча – Европа и Центральная Азия

25-29 октября 2021 года

i. Обработка и архивирование данных

ii. Безопасный доступ к микроданным

Техническая сессия 7: Обработка и архивирование данных, безопасный доступ к микроданным

Елуа Уэдраого
Специалист по статистике
Команда сельскохозяйственных переписей и обследований
Отдел статистики ФАО

СОДЕРЖАНИЕ

Глава 21: Обработка данных

1. Введение
2. Программное и аппаратное обеспечение
3. Тестирование компьютерных программ
4. Виды деятельности по обработке данных
5. Редактирование данных
6. Импутация
7. Валидация данных и составление таблиц

Глава 21: Архивирование данных

1. Введение
2. Передовой опыт обеспечения целостности цифрового контента
3. Оценка и контроль в области сохранения данных

Глава 22: Безопасный доступ к микроданным

Что такое микроданные и зачем они нужны?

1. Метаданные и раскрытие статистической информации
2. Типы доступа

Обработка данных

Введение

- Обработка данных включает в себя кодирование, ввод, редактирование, импутацию, валидацию и табулирование данных.

- Обработка данных зависит от потенциала страны в области информационно-коммуникационных технологий (ИКТ) (т. е. аппаратного и программного обеспечения и инфраструктуры, включая выбор метода сбора данных (например, PAPI или CAPI).

Аппаратное и программное обеспечение

Стратегия в области ИКТ должна быть частью общей стратегии сельскохозяйственной переписи; она в значительной степени зависит от выбранного метода сбора данных и способа проведения переписи. Решение должно быть принято на ранней стадии, чтобы выделить достаточно времени для тестирования и внедрения системы обработки данных.

Необходимо рассмотреть следующие ключевые вопросы управления:

- Стратегические решения для программы переписи, часто связанные со своевременностью и затратами;
- Существующая технологическая инфраструктура;
- Уровень технической поддержки;
- Потенциал сотрудников ведомства по переписи;
- Технологии, использованные в ходе предыдущих переписей;
- Обеспечение эффективности технологий;
- Экономическая целесообразность.

Аппаратное и программное обеспечение – продолжение

Требования к аппаратному обеспечению

Основные характеристики обработки данных сельскохозяйственной переписи:

- Ввод больших объемов данных в сжатые сроки с многопользовательским доступом и режимом параллельной обработки данных на серверах;
- Потребность в больших объемах хранения данных;
- Относительно простые операции;
- Относительное большое количество таблиц, которые необходимо подготовить;
- Широкое одновременное использование файлов необработанных данных;
- Метод ввода данных, выбранный центральным офисом переписи.

Основное аппаратное оборудование:

- Многочисленные устройства сбора данных (ПК, ноутбуки, портативные устройства);
- Центральный процессор/сервер и сети;
- Быстрые графические принтеры с высоким разрешением.

Программное обеспечение

- Позволяет повысить эффективность обработки данных;
- Предпочтительнее использовать стандартное программное обеспечение, поддерживаемое производителем (для обеспечения переноса данных).

Тестирование компьютерных программ

- Для написания компьютерных программ требуется значительное время.
- Компьютерные программы должны тестироваться путем проверки результатов как выявления ошибок, так и составления таблиц для группы из 100-500 вопросников (или меньше, при отсутствии квалифицированного персонала).
- Данные, используемые для таких тестов, должны быть сведены в таблицу вручную для проверки каждого признака или его классификации в таблицах.
- Пробные переписи предоставляют хорошую возможность для окончательного и всеобъемлющего тестирования компьютерных систем и программ, в том числе в отношении передачи данных (в случае использования CAPI, CATI, CAWI).

Виды деятельности по обработке данных

Основные виды деятельности по обработке данных:

1. Кодирование и ввод данных
2. Редактирование данных
3. Импутация
4. Валидация и табулирование
5. Расчет погрешности выборки и дополнительный анализ данных.

1. Кодирование и ввод данных

Кодирование данных: операция, в ходе которой первоначальная информация записанная счетчиками в бумажном вопроснике, заменяется числовым кодом, необходимым для обработки:

- Ручное - Компьютерное

Методы ввода данных:

- Ручной ввод данных;
- Оптическое сканирование;
  - интеллектуальное распознавание символов (ИРС) - оптическое распознавание меток (ОРМ)
- Портативные устройства;
- Веб- и телефонные интервью с использованием компьютера (CAWI и CATI).

Виды деятельности по обработке данных (продолжение)

Ручной ввод данных

- Трудоемкая операция
- Подвержен человеческой ошибке
- Требует большего количества сотрудников
- Требуется строгая процедура проверки
- Простое программное обеспечение

Виды деятельности по обработке данных – продолжение

Ввод данных с помощью портативных устройств

- Метод CAPI с использованием электронных вопросников, в котором ввод данных выполняется непосредственно счетчиками:
  - экономичный
  - позволяет автоматическое кодирование и редактирование
  - имеет функции шаблонов пропуска
  - Требуется тщательное тестирование приложения для ввода данных:
    - Функциональное тестирование
    - Тестирование в использовании
    - Тестирование передачи данных.

Ввод данных в рамках CAWI и CATI

- обычно осуществляется в сочетании с другими методами
- аналогичен вводу данных с помощью портативных устройств - вопросник в режиме онлайн обычно не является точной загружаемой версией бумажного вопросника; скорее, это приложение, которое направляет респондента через вопросник
- тестирование потока вопросов и шаблонов пропуска в интерактивном вопроснике имеет важное значение.

Редактирование данных

- Процесс проверки и корректировки собранных данных переписи.
- Цель: контроль качества собранных данных.
- Редактирование вопросников проводится для:
  - обеспечения согласованности данных и согласованности таблиц (между таблицами и внутри таблиц);
  - для выявления и проверки, исправления или устранения резко отклоняющихся значений.

Ручное редактирование данных (при использовании PAPI)

- Проверка полноты заполнения вопросника для сведения к минимуму числа неполученных ответов.
- Следует начинать как можно скорее после сбора данных и как можно ближе к источнику данных.
- Очень часто ошибки возникают из-за неразборчивого почерка.
- Имеет некоторые преимущества: выявляет бумажные вопросники, которые должны были быть возвращены для заполнения; помогает обнаружить некачественную регистрацию.

Редактирование данных– продолжение

Автоматическое редактирование данных

- Электронное исправление цифровых данных.
- Эффективный подход к редактированию данных переписи с точки зрения затрат, требуемых ресурсов и времени, затрачиваемого на обработку.
- Проверка общей достоверности цифровых данных в отношении следующего:
  - *Отсутствующие данные;*
  - *Установленные допустимые пределы;*
  - *логическая и/или численная согласованность.*
- Два способа:
  - в интерактивном режиме на этапе ввода данных: сообщения об ошибках могут немедленно выводиться на экран и/или отклонять данные, если они не будут исправлены; целесообразен в случае простых ошибок, таких как ошибки, допущенные при неправильном нажатии клавиши, но может значительно замедлить процесс ввода данных. Используется в рамках методов сбора данных CAPI, CATI и CAWI.
  - с использованием пакетной обработки данных: происходит после ввода данных и состоит из проверки множества вопросников в одном пакете. Результатом обычно является файл с сообщениями об ошибках. Используется в рамках всех методов сбора данных.

Автоматическое редактирование данных – продолжение

Две категории ошибок:

- **критические** - должны быть исправлены, поскольку они могут даже заблокировать дальнейшую обработку или ввод данных.
- **некритические** - приводят к недействительным или противоречивым результатам, не прерывая поток последующих этапов обработки. Следует исправить как можно большее количество некритических ошибок, избегая при этом чрезмерного редактирования.

- **Редактирование данных и выявление ошибок может применяться на нескольких уровнях**:
  - На уровне признаков, который обычно называется «проверка допустимых пределов»;
  - На уровне вопросника (проверки осуществляются по соответствующим признакам вопросника);
  - Иерархическое редактирование, которое включает в себя проверку признаков в соответствующих подвопросниках.

3. Импутация

- Процесс решения проблемы недостающих, неправильных или нелогичных ответов, выявленных в ходе редактирования, на основе имеющихся знаний.
- Всякий раз, когда используется импутация, следует делать пометку.
- Обычно используются два метода импутации:

  (a) метод «колд-дек» (статические справочные таблицы)

  (b) метод «хот-дек» (динамические справочные таблицы)
- Может быть произведена вручную или автоматически компьютером:
- При использовании автоматического редактирования и импутации следует учитывать следующие аспекты:
  - непосредственной целью сельскохозяйственной переписи является сбор данных высокого качества. Если обнаружено лишь несколько ошибок, то любой метод их исправления может считаться удовлетворительным;
  - важно вести учет количества обнаруженных ошибок и корректирующих действий (по виду исправления);
  - неполученные ответы всегда можно включать в таблицу в качестве отдельной колонки.
  - дополнительная информация, собранная в переписном листе, полезна для выявления ошибок в ответах.

4. Валидация данных и составление таблиц

4a. Валидация

- Должна выполняться параллельно с другими процессами.
- Все признаки данных должны быть проверены на предмет согласованности и точности по всем категориям на различных уровнях географического агрегирования.
- Макроредактирование – процесс проверки агрегированных данных. Основное внимание уделяется ошибкам, влияющим на публикуемые данные.
- Валидация данных перед тем, как они покинут центр обработки, обеспечивает возможность исправления существенных ошибок в окончательном файле.
- Этот окончательный файл может затем использоваться в качестве исходной базы данных для производства **всех** продуктов распространения.

4b. Составление таблиц

- Крайне важный компонент переписи; являются наиболее наглядными и используемыми результатами переписи.

5. Расчет ошибки выборки и дополнительный анализ данных

- При использовании выборки (например, для дополнительных модулей в рамках модульного подхода или чередующихся модулей в рамках интегрированной программы переписи и обследований) выборочные веса должны рассчитываться и применяться в соответствии с дизайном выборки.
- Данные могут быть агрегированы с использованием формул оценки и соответствующих выборочных весов.
- Данные не могут быть правильно использованы и оценены, если отсутствует указание на ошибку выборки, с которой связаны полученные значения.

Опыт стран: Сертификация валидированных данных в Канаде

- Статистическая служба Канады учредила Комитет по сертификации (в состав которого входят руководители переписи и эксперты в области сельского хозяйства), который рассматривает и официально удостоверяет результаты валидации данных. Каждая переменная переписи рассматривается и удостоверяется по географическим районам. Информация, представляемая Комитету по сертификации, должна:
  - Предвосхищать результаты переписи (прогноз, другие обследования, консультации с отраслевыми экспертами).
  - Согласовывать результаты с текущим социально-экономическим контекстом.
  - Сравнивать результаты с историческими данными, административными данными, данными обследований и другими коррелированными переменными.
  - Описывать влияние обработки и валидации на исходные данные.
  - Рекомендовать Комитету: •опубликовать данные; •опубликовать данные с предостережением; •отложить для дальнейшего изучения до публикации; или •не публиковать данные.

АРХИВИРОВАНИЕ ДАННЫХ

(RLG – OCLC Report, RLG, Mountain View, CA May 2002)

Введение

- Архивирование материалов переписи охватывает множество аспектов, включая техническую документацию, файлы данных, ИТ-программы и т. д.
- Основное внимание уделяется архивированию файлов микроданных переписи.
- Как и другие данные, данные переписи могут иметь культурную и институциональную ценность в будущем. Поэтому важно принимать надлежащие меры для физической защиты данных.
- Следует постоянно подчеркивать необходимость обеспечения резервных копий данных (как в режиме онлайн, так и в режиме офлайн).
- Архивирование данных должно быть полностью включено в процесс планирования переписи и составления бюджета, с тем чтобы своевременно принимать надлежащие меры.
- К счастью, стандарты сохранения цифровых данных позволяют офисам переписи управлять цифровыми данными в долгосрочной перспективе.

Передовой опыт обеспечения целостности цифрового контента (Международная сеть обследования домашних хозяйств, IHSN, 2009)

Аспекты целостности	Соответствующий передовой опыт
Содержание: обеспечение сохранения основных элементов цифрового контента	Офис сельскохозяйственной переписи должен четко определить данные, подлежащие сохранению, и активно управлять ими.
Контроль неизменности: требование регистрации всех изменений в контенте, в идеале начиная с момента его создания	Это требование можно выполнить путем рутинного использования контрольной суммы для обнаружения преднамеренных или непреднамеренных изменений в данных и уведомления управляющих данными с целью принятия мер.
Идентифицируемость: обеспечение уникальной и специфической идентифицируемости контента по отношению к другому контенту во времени	К примеру, офис сельскохозяйственной переписи должен применять и поддерживать постоянный идентификационный подход, представляющий собой систему присвоения и управления устойчивыми идентификаторами, которая позволяет последовательно и однозначно ссылаться на цифровые материалы с течением времени.
Происхождение: цифровой контент можно отследить вплоть до его источника (момента создания) или, как минимум, от момента помещения на хранение в надежное цифровое хранилище.	Офис сельскохозяйственной переписи должен регистрировать информацию (собранную в качестве метаданных) о создании и действиях, которые повлияли на контент с момента его создания (например, помещение данных на хранение в архив, перенос данных из одного формата файла в другой).
Контекст: документирование и управление взаимосвязями цифрового контента	Офис сельскохозяйственной переписи, сохраняющий данные, должен документировать взаимосвязи в рамках собственного цифрового контента и по возможности по отношению к данным, управляемым другими организациями

Оценка и контроль в области сохранения данных

- **Проведение собственной оценки деятельности по сохранению данных** помогает офисам по проведению переписи в соблюдении стандартов цифрового сообщества и в «рассмотрении их цифровых активов с точки зрения охвата, приоритетов, ресурсов и общей готовности решать проблемы сохранения цифровых данных».
- В качестве следующего шага офисы переписи могут разработать собственную политику сохранения данных, отражающую «мандат организации по сохранению данных».
- Рекомендуются официальные инструменты оценки, такие как:
  - Data Seal of Approval – краткий набор руководящих принципов, которыми офисы переписи могут воспользоваться для самостоятельного проведения оценки.
  - Контрольный список надежного цифрового хранилища – строгий стандарт ИСО (16363);
  - Метод оценки цифрового хранилища на основе оценки риска (DRAMBORA).
- *«Десять принципов для хранилищ цифровых данных»* обобщают основные критерии, которым должны соответствовать надежные хранилища.

БЕЗОПАСНЫЙ ДОСТУП К МИКРОДАННЫМ

(RLG – OCLC Report, RLG, Mountain View, CA May 2002)

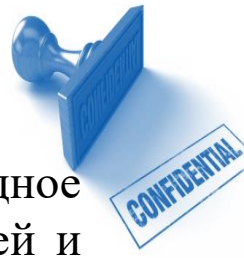
Что такое микроданные и зачем они нужны?

- Микроданные – это информация, записанная респондентом или получаемая от него при проведении обследования или переписи.
- В рамках сельскохозяйственной переписи это касается данных, собранных об аграрных хозяйствах.
- Каждая строка микроданных соответствует хозяйству, а каждый столбец – переменным данных.
- Помимо значений переменных исследователям и другим пользователям данных необходимы метаданные, которые помогают им понять коды, определения и понятия, лежащие в основе собранных данных: *что они измеряют и как они были созданы, чтобы понять качество данных*.
- Микроданные позволяют исследователям использовать данные переписи или обследования для изучения вопросов, требующих более детального анализа, чем первоначальные таблицы.

Сельскохозяйственные микроданные

- Одна из **особенностей** сельскохозяйственных данных заключается в том, что аграрные хозяйства, поскольку они являются единицами производства, подпадают под определение коммерческих предприятий.
- Риски и методы контроля за раскрытием деловой информации отличаются от тех, которые используются в обследованиях домохозяйств: данные о крупных коммерческих фермах часто представляют собой небольшие целевые совокупности, в результате чего их труднее обезличить.
- Вместе с тем, характеристики данных сельскохозяйственных переписей и обследований имеют общие черты с характеристиками данных как *обследований предприятий*, так и *обследований домохозяйств*.

Конфиденциальность и правовые рамки



Конфиденциальность

- При выборе системы доступа к микроданным первоочередное внимание уделяется удовлетворению потребностей исследователей и обеспечению максимальной защиты конфиденциальности респондентов.
- Предоставление доступа к микроданным требует, чтобы статистические учреждения обеспечивали баланс между спросом со стороны исследовательского сообщества и законодательными требованиями в отношении сохранения конфиденциальности информации.

Правовые и политические рамки

Статистическое учреждение должно соблюдать правовые рамки и уставы, в соответствии с которыми оно осуществляет свою деятельность: крайне важно заручиться поддержкой респондентов и в то же время подтвердить, что существуют методы обеспечения конфиденциальности и предоставления микроданных для статистических целей.

- Важно разработать четкую политику в отношении мер, которые могут быть приняты учреждением в отношении доступа к микроданным переписи, и обеспечить прозрачность предоставления этой информация общественности.

Раскрытие статистической информации

- Обеспечение конфиденциальности означает направление всех усилий на гарантирование того, чтобы файл не раскрывал данные, позволяющие идентификацию.
- **Раскрытие** происходит, когда пользователь файла микроданных обнаруживает или узнает ранее неизвестную ему информацию о респонденте переписи или обследования, или когда существует вероятность того, что аграрное хозяйство или лицо в домохозяйстве владельца будет реидентифицировано пользователем микроданных с использованием информации, содержащейся в файле.
- Раскрытие может происходить двумя основными способами:
  - **Раскрытие личности:** происходит, когда в файле остается прямой идентификатор (например, имя и фамилия, номер телефона или адрес), с помощью которого можно установить личность респондента.
  - **Раскрытие атрибутов:** происходит, если какой-либо атрибут или сочетание атрибутов (например, крупная коммерческая ферма или тип редкой сельскохозяйственной культуры) могут быть непосредственно связаны с конкретным респондентом. Лица, обладающие знаниями о регионе смогут идентифицировать этого респондента.
  - **Остаточное раскрытие** происходит, когда последовательные извлечения информации из файла можно сопоставить (вычесть), чтобы изолировать то или иное значение респондента.

Метаданные и раскрытие статистической информации

Контроль за раскрытием статистических данных

Под контролем за раскрытием статистических данных понимается процесс обеспечения соблюдения требований конфиденциальности, регулирующих работу Национальной статистической службы (НСС), и минимального риска раскрытия информации о респонденте (например, путем обезличивания данных).

- Риск раскрытия информации зависит от многих факторов, таких как:
  - чувствительность данных;
  - наличие внешних источников информации, которые могут быть использованы для идентификации респондентов путем сопоставления комбинации переменных, способных идентифицировать респондента;
  - возможность объединения предоставленных данных с данными из других общедоступных источников;
  - является ли файл микроданных результатом выборочного обследования или переписи, проведенной методом сплошной регистрации.

DISCLOSURE

Контроль за раскрытием статистических данных

Некоторые процедуры контроля за раскрытием статистической информации:

- **Удаление прямых идентификаторов**, таких как имена и фамилии, адреса, номера телефонов, подробности расположения аграрных хозяйств.
- **Удаление косвенных идентификаторов**, таких как подробности расположения единиц наблюдения. Сюда входят географические координаты, расположение сегментов выборки, местоположения участков или сегментов, записанные либо в качестве атрибутивной информации, либо в качестве элементов территориальной выборки.
- **Применение методов обезличивания** данных на основе выявленного риска раскрытия. Подробные технические детали методов и программного обеспечения, используемых для обезличивания, не обсуждаются в данном разделе, но широко доступны в литературе.
- **Проверка файла** на утрату полезности и информации.

DISCLOSURE

Типы доступа к микроданным

- **Файлы для открытого пользования:** проходят процесс строгого контроля раскрываемой статистической информации с тем, чтобы минимизировать возможность идентификации респондентов. Исследователям требуется принять определенные условия пользования данными, оговоренные в электронном соглашении, «подтверждаемом щелчком мыши» («click-through»).
- **Лицензированные файлы:** также обезличены, но с возможностью применения меньшего количества процедур контроля раскрываемой статистической информации в зависимости от характера файла и политики производителя.
- **Средства удаленного доступа:** дает возможность исследователям предоставлять алгоритм, который они будут использовать в своем анализе (с помощью инструментов SAS, SPSS, STATA или R) вместе с искусственным файлом, дублирующим структуру и содержание наборов фактических данных. Центральный офис переписи запускает алгоритм на основе набора фактических данных и проверяет результаты на раскрытие конфиденциальных данных перед возвратом пользователю.
- **Анклавы данных:** помещение в пределах статистической организации, куда исследователи могут прийти для проведения своих исследований с использованием подробных файлов. Эти файлы являются самыми детальными (после основного файла) файлами, доступными для исследователей.
- **Условный сотрудник:** прием исследователя/пользователя на работу в учреждение в качестве временного сотрудника.

СПАСИБО ЗА ВНИМАНИЕ