



## Экономический и социальный совет

Распространение: Общее  
18 апреля 2012 г.

Оригинал: английский

---

### Европейская экономическая комиссия

#### Конференция статистиков Европы

#### Заседание экспертной группы ЕЭК ООН - Евростат по проведению переписи с использованием регистров

Женева, 22-23 мая 2012 г.

Сессия 1: Опыт использования регистров в переписи

### Модель преобразования административных данных в статистические данные

#### Записка Центрального статистического офиса Польши<sup>1</sup>

#### I. Введение

1. Преобразование наборов данных из административных источников в статистические данные нацелено на улучшение качества данных из внешних источников и повышение их применимости в статистике.
2. Цель работы с административными данными заключается в получении достаточно полного набора данных в части субъективной и объективной полноты, соответствующей классификационным стандартам, определениям и основным категориям и, таким образом, в эффективном использовании административных данных.
3. Преобразование включает в себя загрузку исходных данных в операционную среду, их последующий контроль и корректировку. Используя решения, знания и опыт, приобретенные за время работы над сельскохозяйственной переписью 2010 г. и переписью населения и жилого фонда 2011 г. в Польше, была создана модель преобразования совокупности данных из административных источников в статистический регистр.
4. Преобразование наборов данных из административных источников в статистические данные в пределах идентификации и адреса включает: личный идентификатор (PESEL), налоговый идентификатор (NIP), регистр хозяйствующих

---

<sup>1</sup> Автор - Pawel Murawski.

субъектов (REGON), имена, фамилии, адресные данные – страна, воеводство (провинция), регионы, муниципалитеты, города, улицы и номера недвижимости. Очистка идентификаторов PESEL, NIP, REGON требует подтверждения длины идентификаторов и проверочный символ.

5. Полученные данные могут быть, как и плохого, так и очень хорошего качества, поэтому возможность использования регистра в статистических целях зависит от оценки. Анализ качества исходных данных из административных записей должен быть основан на показателях качества.

- Показатели качества административных регистров:
  - своевременность данных
  - методологическая совместимость
  - полнота
  - идентификационные стандарты, используемые в регистре
  - полезность
  - совместимость данных из административных источников с данными, полученными в ходе обследования.
- Показатели качества в процессе обработки регистров данных:
  - коэффициент ошибок избыточного охвата
  - коэффициент ошибок неполного охвата – субъективный показатель полноты
  - объективный показатель полноты
  - коэффициент вменения
  - индекс корректировки данных
  - индекс интеграции данных из различных источников.

## II. Этапы преобразования регистров

6. Процесс преобразования административных данных в статистические данные начинается с импортирования данных в операционную среду, где путем соответствующих механизмов наборы данных преобразуются, а затем интегрируются. Статистический регистр, подготовленный таким образом, отправляется в аналитическую среду.

7. Первый этап работы с набором данных заключается в их извлечении и загрузки в операционную среду, используя программное обеспечение Института SAS.

8. Формат полученных данных бывает разным, например, .txt, .xls, .csv, .xml, базы данных MS SQL. Импортирование означает объединение данных из различных исходных систем и преобразование данных в удобный для обработки формат – таблицы SAS.

9. Проверка правильности данных и их структуры является неотъемлемой частью импорта данных. Это включает, в частности, количество импортированных записей (соответствует ли количеству записей, представленных провайдером

данных) и подтверждение правильности распределения данных в отдельные столбцы (что подразумевает проверку того, содержится ли текст в столбцах с текстовым значением, подходит ли длина поля для переменных данных и т.д.).

10. Процесс извлечения заключается в загрузке не всех, а только необходимых и отобранных данных. В связи с этим во время подготовки к Сельскохозяйственной переписи 2010 г. и Национальной переписи населения 2011 г. была проведена методологическая работа по диагностированию содержания и качества государственных регистров. В конечном счете, для применения в переписи было отобрано и получено 29 источников из государственных информационных систем.

11. Преобразование данных подразумевает серию действий в операционной среде, в которые входят: профилирование – создание отчета о качестве данных, унификация данных, разбор (разделение) или объединение переменных, приведение в соответствие со схемой, перенос, проверка достоверности, исключение избыточных данных, интеграция данных.

12. Преобразование данных – профилирование: после успешного извлечения набора данных, происходит процесс, который называется профилированием. Мы создаем отчет о качестве данных с тем, чтобы можно было проверить (на уровне цифр и процентов) коэффициент ошибок для каждой переменной в наборе данных. При профилировании мы можем получить информацию о числе завершенных записей, о количестве уникальных вводов, об образцах и некорректных данных.

13. Преобразование данных – стандартизация и разбор: стандартизация данных - унификация и приведение к определенному стандарту – величины, возникающие в определенных столбцах. Разбор – это разделение переменных, например, разделение одного столбца «адрес» на столбцы «улица», «город», «номер дома» или частичное имя и фамилия из одного текстового поля.

14. Преобразование данных – конверсия – это также является процессом преобразования. Конверсия подразумевает замену величин, сохраненных по-разному для одной и той же информации, кодирование информации.

15. Преобразование данных – подтверждение действительности - это процесс проверки корректности данных и корректировки аномальных значений согласно алгоритмам, подготовленным методистами. Иногда необходимо также исключить записи, улучшение которых не является возможным, из последующей обработки. Посредством этого процесса мы можем получить данные более высокого качества. Подтверждение действительности проводится на наборах данных, которые предварительно были «очищены» на предыдущих этапах работ.

16. Преобразование данных – исключение избыточных данных - это процесс удаления повторяющихся единиц и соединения информации в одних и тех же записях. Для этого требуется детальный анализ, который часто включает анализ правовых актов. Такой анализ индивидуален для каждого регистра. В результате исключения избыточных данных мы получаем одну уникальную запись всей возможной и уникальной информации.

17. Преобразование данных – интеграция данных - это процесс выбора самого лучшего, самого актуального и правильного значения из нескольких регистров. Это процесс создания статистической записи, которая будет доступна для использования аналитиками.

18. Статистический регистр - это передача с операционного участка и аналитическую среду. В данном процессе важно использовать механизмы для быстрой загрузки большого объема данных. В аналитической части продолжаются

дальнейшие работы по данным, такие как создание сводных таблиц или составление отчетов.

## **V. Заключение**

19. Процесс преобразования данных в статистические данные является сложным процессом, состоящим из нескольких этапов, которые меняются в зависимости от вида регистра. Работа в этих рамках требует предварительной подготовки основы и методологии. Правильное преобразование данных также необходимо для проведения тщательного анализа. Необходимо определить разные случаи.

20. Данный процесс является важным, если мы хотим использовать данные из административных регистров. Проводимый перевод государственного управления в цифровую форму, расширение существующих и создание новых баз данных делает возможным более широкое использование такого вида данных в статистической работе.

---