

Распространение: Общее
18 апреля 2012 г.

Оригинал: английский

Европейская экономическая комиссия
Конференция европейских статистиков

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

Женева, 22-23 мая 2012 года

Сессия 3: Наличие, полнота и качество данных регистров и других источников

Регистр адресов и зданий – комбинация разных регистров

Записка Федерального статистического управления Германии¹

Резюме

Наиболее сложной задачей в рамках переписи 2011 г. в Германии является комбинация разных компонентов обследований с тем, чтобы можно было синтезировать однозначные записи данных, содержащие соответствующие переменные переписи. Поскольку источники данных являются разнородными – это регистры, полный подсчет и выборочное обследование – необходимо обеспечить связь между компонентами переписи. Для этого был создан базовый регистр, содержащий все адреса с жилой площадью и заселенными жилыми помещениями в Германии. Далее описываются различные источники данных, которые были применены для построения этого базового регистра. Затем в документе описывается предварительная обработка данных, включая гармонизацию и стандартизацию, и процесс проверки данных.

¹ Автор - Susanne Stiglmayr.

I. Модель переписи 2011 г. и основной регистр

1. 9 мая 2011 г. в Германии началась первая за более чем 20 лет перепись населения. В настоящее время данные по населению основываются на обновленных результатах последних переписей населения и, соответственно, становятся все более неточными. Если в 1987 г. в Федеративной Республике Германии и в 1981 г. в Германской Демократической Республике, соответственно, проводилась полная перепись всех граждан (традиционная перепись), то перепись 2011 г. будет также полной, но на основе регистров. Причина изменения методологии связана с целью по снижению нагрузки на население. Использование административного регистра позволяет оценить все данные без необходимости связываться с самими гражданами. Кроме того, новый метод переписи 2011 г. позволяет сократить затраты. С другой стороны, для обработки и оценки этих огромных объемов данных необходимы высокоразвитые информационные технологии.

2. При традиционной переписи имеется возможность включить всю информацию, имеющую отношение к переписи в один переписной лист и, соответственно, в один массив данных. При использовании же в Германии метода на основе регистров информация имеется в разных массивах данных, которые собираются отдельно и которые не имеют стандартного идентификационного номера для граждан или адресов. Поскольку источники данных являются разнородными – это регистры, полный подсчет и выборочное обследование – необходимо обеспечить связь между компонентами обследования. Для этого был создан базовый регистр, содержащий все адреса с жилой площадью и заселенными жилыми помещениями в Германии. Данный регистр адресов и зданий (Anschriften- und Gebäuderegister, AGR) функционирует в качестве связующего элемента в модели переписи и, таким образом, обеспечивает возможность привязки переменных переписи. Функции AGR детально описаны в подразделе 2 Раздела 2 Закона о подготовке к переписи 2011 г.:

Регистр адресов и зданий служит

- 1) *в целях управления процессами, относящимися к переписи зданий и жилищного фонда, и управления потоком всех обследований первичных данных, составляющих часть всеобщей переписи,*
- 2) *в целях подготовки выборочных обследований, которые будут проводиться в рамках переписи, и будет служить в качестве основы выборки,*
- 3) *в целях координации обследований в рамках переписи, в целях комбинации данных из различных источников на этапе проведения переписи и в целях проверки сплошного охвата переписью всех зданий, жилищ и граждан,*
- 4) *в целях разработки системы пространственного анализа и представления статистических результатов и создания основы для оценки результатов переписи в малых регионах.*

3. Для того, чтобы AGR мог выполнять свою ведущую роль по управлению переписью 2011 г., AGR должен быть расширен и использоваться уже на этапе подготовки. Для этого были получены различные имеющиеся административные регистры, данные были отредактированы, привязаны на уровне адресов в стандартизированной форме и агрегированы в единый массив данных. Последний использовался в качестве основного регистра для целей контроля и координации на этапе подготовки обследования. AGR содержит все адреса жилых площадей и заселенных жилых помещений; это означает все адреса, где проживают люди, и

здания, имеющие отношение к переписи. Вместе с соответствующими адресами хранятся различные характеристики и переменные, которые содержат дополнительную информацию об адресе.

II. Создание регистра адресов и зданий

4. Для переписи 2011 г. в Германии необходим базовый регистр, который объединяет различные источники данных. В связи с отсутствием единого для всех регистров идентификационного номера в переписи используется адрес для того, чтобы обеспечить привязку между источниками данных. В связи с этим основной файл переписи – регистр адресов и зданий – должен содержать все адреса жилых помещений. К сожалению, такой регистр отсутствует в Германии и должен быть создан на этапе подготовки к переписи. Для создания AGR было использовано три центральных регистра официальных статистических данных. Источники данных и условия определены в Разделе 2 Закона о подготовке к переписи 2011 г. AGR основывается на источниках данных регистров населения (MR), Федерального агентства занятости (BA) и адресные данные Федерации (GAB) с привязкой к географическим координатам. Эти массивы данных были объединены в целях обеспечения полного охвата адресов, необходимых для переписи.

5. Администраторами регистров населения (MR) являются органы регистрации граждан. В Германии каждый муниципалитет хранит демографические и географические сведения о гражданах, проживающих в муниципалитете. Дальнейшие разъяснения можно найти во вспомогательном материале к данному заседанию «Регистры населения в Германии – основной источник данных в переписи 2011 года». Для сбора всех адресов жилых помещений важной является такая характеристика регистров населения как «место обычного проживания».

6. Регистр Федерального агентства занятости (BA) содержит не только адресные переменные, но также переменные, связанные с занятостью, такие как вид деятельности, род деятельности и место работы. Каждый работающий в Германии регистрируется в регистре Федерального агентства занятости, при этом указываются его личные данные, такие как имя, дата рождения и адрес.

7. Третьим источником для создания AGR являются адресные данные Федерации с привязкой к географическим координатам (GAB). В отличие от первых двух источников адресные данные Федерации с привязкой к географическим координатам не содержат какие-либо личные данные. Регистр включает все участки земли с их географическими координатами. Это означает, что каждый адрес здания, где органы земельной регистрации Земли (Länder) проводили изыскания, включен и объединен с географическими координатами. Файл включает как заселенные, так и незаселенные жилые помещения.

8. Поскольку AGR является агрегированным массивом данных на адресном уровне, регистры, используемые для его создания, должны иметь одинаковый уровень агрегирования и одинаковое определение адреса. Данные Федерального агентства занятости и регистров населения являются данными по гражданам, т.е. адрес может быть указан несколько раз, иногда даже с разным написанием. В связи с этим все три регистра были, прежде всего, агрегированы на уровне адресов. Следующим шагом необходимо было гармонизировать определение адреса. Каждый из используемых регистров имеет собственное толкование адреса. С другой стороны, это является следствием предназначения регистров. Изначально регистры не ведутся для того, чтобы представлять действительный адрес. С другой стороны, федеральная структура Германии поддерживает разнородность массивов данных. Однако

децентрализованное хранение регистров приводит к сложным проблемам обеспечения качества данных. Между регистрами и отдельными органами, осуществляющими сбор данных в Землях и муниципалитетах, отсутствует стандартизация данных, поскольку регистры не создавались для последующего объединения на федеральном уровне. Как следствие, данные по адресным переменным представляются по-разному, что иногда обусловлено личными предпочтениями ответственного персонала и различиями в процессах по обновлению данных. Например, вся информация по адресам в регистре Федерального агентства занятости хранится только в одной переменной и в различных формах. Мы определили стандартный формат адреса в регистре адресов и зданий следующими характеристиками:

1. *уровень улицы*
 - a) *официальный код муниципалитета*
 - b) *почтовый индекс*
 - c) *подмуниципалитет*
 - d) *улица*
2. *уровень адреса*
 - a) *номер дома*
 - b) *дополнение к номеру дома.*

9. Исходя из этого определения адреса каждому уникальному набору данных присваивается собственный идентификационный номер (идентификатор адреса). Этот идентификационный адрес используется в течение всей переписи во всех частях и массивах данных переписи.

10. В целях гармонизации данных и обеспечения качества AGR в массивах данных GAB, MR и BA были предприняты различные меры по обработке данных. Наиболее важными мерами были структурный анализ и стандартизация адресных переменных и привязка переменных, т.е. сопоставление названий улиц с внешними источниками. Здесь структурный анализ означает, что адресные данные конкретных регистров были разложены на адресные компоненты. Поскольку затем в разных регистрах появились одни и те же адресные компоненты в стандартной форме, было выполнено основное требование по автоматизированной сравнимости в нескольких источниках. Таким образом, адреса в BA, которые были представлены в одной колонке, могли быть разложены в стандартном формате адреса. Для этого был применен ряд правил, которые, например, в простейшей форме, предусматривают написание названий всех улиц заглавными буквами. После структурного анализа и стандартизации данные в каждом массиве данных были агрегированы на уровне улиц и объединены с другим массивом данных через муниципалитет и название улицы. Цель этого действия заключалась в том, чтобы удостовериться в правильности и существовании названий улиц, если они имелись в идентичной форме, как минимум, в двух источниках данных. Те улицы, которые было невозможно объединить в автоматизированном режиме, передавались в статистические органы Земель для их корректирования на экране. После возвращения скорректированные данные интегрировались в массив данных в автоматизированном режиме, что позволило существенно повысить качество массива данных. Поскольку на каждой улице имеется несколько адресов, такая двухэтапная процедура – с отдельной обработкой на уровне улицы – может быть применена для корректирования многих адресов. В случае когда источником ошибки является просто неправильное название улицы, исправление названия улицы в идеале позволяет объединить несколько адресов,

которые кажутся разными, в один адрес. Таким образом, исправление названия улицы может решить ряд проблем с адресом. Помимо соотнесения названий улиц было необходимо также гармонизировать ряд других переменных. Несмотря на то, что идентификаторы адреса, такие как официальный код муниципалитета или почтовый индекс может измениться с течением времени, эти переменные необходимо было унифицировать. В разрезе одного источника данных или в разрезе разных источников данных официальный код муниципалитета или почтовый индекс может происходить из разных территориальных классификаций. Также они могут меняться с течением времени в ходе переписи. В обеих комбинациях идентификаторы адресов необходимо было унифицировать на базовую дату переписи 2011 г.

11. По крайней мере, три подготовленных источника данных - MR, VA и GAB – были объединены в одну базовую таблицу. Новая база данных содержала все адреса из трех регистров. Основной базой данных является идентификатор адреса с адресными переменными, включая географические координаты. В AGR можно увидеть источник адреса (MR, VA и/или GAB).

III. Контроль качества AGR

12. Таким образом, будучи ведущим файлом, AGR занимает центральное место в модели переписи. Унифицированный фонд адресов обеспечивает, что стандартизированные обозначения AGR будут также использоваться в других компонентах обследования и что в силу структуры базы данных в ходе переписи 2011 г. адреса могут поддерживаться только единожды и централизованным образом в AGR.

13. Ряд мер направлен на проверку качества AGR.

14. Использование трех регистров было связано не только с тем, что сведения, актуальные для переписи, должны были быть сведены из разных регистров, но также служило цели контроля качества данных. Поскольку регистры являются отдельными источниками, комбинация данных также могла использоваться в качестве индикатора достоверности адреса. Можно было предположить, что адрес, имеющийся в одинаковой версии в разных файлах, действительно существует.

15. Поскольку перепись охватывает только жилые помещения и здания с жилой площадью, при создании AGR необходимо определить может ли конкретный адрес быть классифицирован как «адрес с жилой площадью». Если жилая площадь отсутствует, такой адрес должен быть исключен из дальнейшей деятельности по переписи. Согласно правилу, что адреса, имеющиеся, как минимум, в двух источниках (MR, VA и GAB), являются адресами с жилой площадью, адреса, имеющиеся только в одном из трех источников, должны были быть проверены местными статистическими органами. Необходимо было проверить является ли адрес адресом с жилой площадью. Такая проверка предусмотрена в Разделе 14 Закона о подготовке к переписи 2011 г. Дальнейший процесс показывает необходимость контролирования характеристики (жилая площадь или нежилая) адреса. В рамках контроля качества критерий касательно того, является ли площадь жилой или нежилой, необходимо было контролировать постоянно, независимо от Раздела 14 Закона о подготовке переписи 2011 г. Цель по наличию или отсутствию жилой площади меняется постепенно в течение этапа подготовки переписи и должна контролироваться для всех адресов с течением времени.

16. С учетом этого имеется специальное приложение для обновления AGR. Центральное приложение дает возможность редактирования адресных переменных в

AGR. Поскольку такие данные как названия улиц могут меняться, постоянное корректирование данных регистра должно быть частью контроля качества. Также может меняться статус по заселенности и незаселенности. Приложение также обеспечивает возможность добавления новых адресов в AGR. Поскольку AGR должен отражать все адреса с жилой площадью и заселенные жилища в Германии, интеграция новых заселенных жилищ и зданий с жилой площадью является еще одним свойством приложения. Федеральное статистическое управление, а также местные статистические органы земель, работают с приложением с целью обновления AGR.

17. Реализуется много процессов по обновлению AGR. В связи с этим, поддержание актуальности и полноты AGR требует интеграции изменений после создания базового фонда. Такое обновление обеспечивается, во-первых, посредством интеграции нескольких корректировок с 2008 по 2011 гг. и, во-вторых, посредством привязки дополнительных компонентов обследований. Данные регистров населения и адресные данные Федерации с привязкой к географическим координатам повторно представляются в Федеральное статистическое управление и интегрируются в запас данных. Пособием получения таким образом новых адресов обеспечивается полнота AGR. Другие выгоды от информации в AGR достигаются посредством привязки других компонентов обследований. AGR будет завершен путем привязки регистра адресов помещений коллективного проживания и институциональных домохозяйств. Благодаря этому новые адреса будут интегрированы в AGR. А посредством привязки регистра переписи зданий и жилищного фонда будет усовершенствован AGR. Привязка собственников зданий ко всем заселенным жилищам и зданиям с жилой площадью преследует две цели: во-первых, перепись зданий и жилищного фонда проверяется на полноту, поскольку собственник должен указываться в каждом адресе, актуальном для переписи, содержащемся в AGR. Во-вторых, адреса из переписи зданий и жилищного фонда могут быть добавлены в AGR, если они существуют и еще не были включены.

18. AGR может быть усовершенствоваться, по крайней мере, на этапе проведения самой переписи. Компоненты переписи обеспечивают обновление фонда адресов путем интеграции результатов обследования. Например, в рамках переписи зданий и жилищного фонда собственник получает переписной лист на определенный адрес и указывает в нем дополнительную собственность. Используя эту информацию, AGR может также обновляться в ходе переписи. Результаты выборочного обследования могут также улучшить AGR. Например, счетчик обнаруживает, что имеются адреса без зданий. Такая информация интегрируется в AGR. Необязательно, что будут выявлены новые адреса, но в основном может быть изменен показатель «заселен» или «не заселен».

IV. Заключение

19. В целях комбинирования разных компонентов переписи 2011 г. в Германии был необходим базовый регистр. Хотя такой регистр не существует в Германии, было принято решение создать основной регистр адресов. Используя этот базовый регистр, все компоненты переписи населения и переписи зданий и жилищного фонда могут быть привязаны друг с другом через адрес. Комбинирование данных из разных источников данных на уровне адресов являлось одной из наиболее сложных задач переписи 2011 г. Опыт, полученный с такого рода комбинированием данных без числовых идентификаторов в рамках переписи 2011 г., позволяет сделать конкретные предложения для дальнейшей проработки новой переписи в Германии. Очевидно, что статистические органы в Германии хотели бы предпринять усилия для того, чтобы

данные по зданиям, жилищному фонду и домохозяйствам, можно было собирать на основе регистров, что будет означать оперативный и недорогостоящий сбор данных. Обширный опыт по созданию АГ поможет нам в достижении этой цели.
