Кодирование письменных ответов на вопросы переписи населения

Выбор тем в международных переписях населения[1]

Выпущено в июне 2020 года

ВВЕДЕНИЕ

Все инструменты переписи населения содержат вопросы, требующие от респондентов ответа в письменном виде. Письменные поля позволяют отвечать в свободной форме без использования готовых категорий ответа. Таким образом, письменные ответы дают респондентам возможность отвечать на вопросы без ограничений, налагаемых заранее подготовленными категориями. Однако наличие письменных полей в анкете создает определенные сложности для национальных статистических служб (NSO, в соответствии с английским акронимом) в ходе сбора и обработки данных.

Данная техническая записка из серии «Избранные темы международных переписей населения» (STIC, в соответствии с английским акронимом) предоставляет службам NSO сведения о международно признанных стандартах автоматизированного кодирования письменных ответов на вопросы переписи населения.

ФОРМАТЫ ВОПРОСА

Большинство вопросов в анкете переписи населения являются закрытыми, то есть предусматривают стереотипные альтернативные ответы. Для таких вопросов приводится фиксированный список вариантов ответа, при этом респонденту, как правило, предлагается выбрать один или более из них. Существуют также вопросы открытого типа — для них не приводятся готовые категории ответа, и респонденты могут отвечать на них в свободной форме своими словами. В бумажных анкетах переписи населения, предназначенных для сканирования, ответы на вопросы открытого типа обычно заносятся в поля с фиксированным максимально допустимым количеством символов.

У вопросов закрытого и открытого типа есть свои преимущества и недостатки. Обычно закрытые вопросы с фиксированным списком ответов легче поддаются анализу, и на них проще отвечать, благодаря чему уменьшается нагрузка на респондентов и NSO в ходе сбора, обработки и анализа данных. Однако в заранее подготовленные категории ответа часто не входит весь спектр ответов, которые могли бы дать респонденты, если бы вопрос был задан в открытом формате. Письменные ответы на открытые вопросы позволяют респонденту выражаться своими словами, используя собственные знания и представления. Однако это создает определенные сложности, так как службе NSO труднее обрабатывать и анализировать ответы на открытые вопросы. Открытые вопросы также являются существенным источником систематической ошибки, поскольку даже идеально составленный вопрос может быть по-разному понят разными респондентами.

Обычно исследователи, работающие в NSO, задействуют открытый вопрос, когда его неудобства перевешиваются ограничениями вопроса закрытого типа. Есть также случаи, когда вопрос открытого типа должен быть задан обязательно — см. пояснения ниже.

[1] Данная техническая записка является одной из серии *«Избранные темы международных переписей населения»,* в которой рассматриваются вопросы, представляющие интерес для международного статистического сообщества. Бюро переписи населения США помогает странам совершенствовать национальные системы статистики, содействуя устойчивому расширению статистических компетенций.

U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

United States Agency for International Development
www.usaid.gov

United Nations Population Fund
www.unfpa.org

Таблица 1.

Экономическая деятельность, род занятий и отрасль: перепись населения и жилищного фонда Ботсваны 2011 года

ВСЕ ЛИЦА В ВОЗРАСТЕ 12 ЛЕТ И СТАРШЕ					
ЭКОНОМИЧЕСКАЯ ДЕЯТЕЛЬНОСТЬ				РОД ЗАНЯТИЙ	ОТРАСЛЬ
Чем ... главным образом занимался(лась) со времени дня независимости 2010 года? **Сезонные работы** 01 Оплачиваемые 02 Неоплачиваемые **Несезонные работы** 03 Оплачиваемые 04 Неоплачиваемые **Другое** 05 Поиск работы 06 Работа по дому 07 Учащийся 08 На пенсии 09 Болезнь Другое (укажите)	Выполнял(а) ли ... какую-либо работу ради оплаты, получения выгоды или пользы для дома как минимум в течение 1 часа за последние 7 дней? 1 Да (ПЕРЕЙТИ К A22) 2 Нет [Если нет, работал(а) ли ... на собственной земле/с собственным скотом?]	Чем занимался(ась) ... с тех пор, как он(а) не работал(а)? 1 Активный поиск работы 2 Работа по дому 3 Учащийся 4 На пенсии 5 Болезнь Другое (укажите) **[Если женского пола, ПЕРЕЙТИ К A25** **Если мужского пола, ПЕРЕЙТИ К следующему лицу]**	Кем работал(а) ... на протяжении последних 7 дней? 1 Сотрудник — оплата денежными средствами 2 Сотрудник — оплата товарами 3 Самозанятый (сотрудники отсутствуют) 4 Самозанятый (имеются сотрудники) 5 Безвозмездная работа на семейном предприятии 6 Работа на собственной земле/пастбище	Какие виды работ ... выполнял(а) за последние 7 дней? Точнее, каковы были главные задачи и обязанности? *(Задавайте наводящие вопросы по мере необходимости, описывайте специальность двумя или более словами)*	Каков был основной вид продукции, услуг или деятельности на месте работы ... ? *(Задавайте наводящие вопросы по мере необходимости, описывайте отрасль двумя или более словами)* **Если женского пола, ПЕРЕЙТИ К A25, иначе** ПЕРЕЙТИ К следующему лицу
A19(2)	A20	A21	A22	A23(3)	A24(4)

Источник: Statistics Botswana, 2011.

Вопросы открытого и смешанного типа

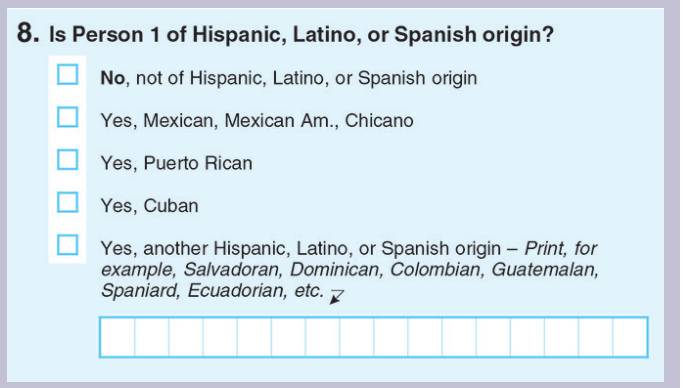
Все анкеты переписи населения содержат как минимум один вопрос, требующий письменного ответа, — имя и фамилия жителей домохозяйства. Кроме того, службы NSO почти всегда дают возможность предоставить письменный ответ на некоторые вопросы. Например, открытыми нередко являются вопросы об отрасли и профессии. Закодировать или перечислить сотни профессий, имеющихся в какой-либо отрасли и списке кодов специальностей, непросто, особенно если анкета бумажная. Поэтому службы NSO нередко разрешают письменные ответы, когда требуется зафиксировать больше подробностей, чем можно было бы отобразить в бумажной или цифровой анкете.

В таблице 1 приведен пример раздела «Трудоустройство» переписи населения Ботсваны 2011 года. Обратите внимание, что вопросы об экономической деятельности представлены в закрытом формате, тогда как вопросы о профессии и отрасли являются открытыми.

На рисунке 1 приведен пример вопроса Переписи населения США 2020 года, где поле для письменного ответа, объединенное с четырьмя клетками для отметки галочкой, предназначено для сбора подробной информации об испаноязычном и латиноамериканском происхождении. Первая клетка для отметки галочкой адресована респондентам, которые не имеют испаноязычного, латиноамериканского или испанского происхождения. Следующие три клетки адресованы самым крупным группам, имеющим происхождение из испаноязычной страны, — «мексиканец (мексиканка)», «пуэрториканец (пуэрториканка)», «кубинец (кубинка)». Последняя клетка для отметки галочкой, «Иное испаноязычное, латиноамериканское или испанское происхождение», сопровождается группой подробных примеров и специальным письменным полем для сбора

Рисунок 1.

Вопрос смешанного формата: форма домохозяйства Переписи населения США 2020 года



Источник: U.S. Census Bureau, 2020 Census.

данных по всем остальным группам испаноязычного и латиноамериканского населения. Такое сочетание клеток для отметки галочкой и письменного поля позволяет всем респондентам сообщить о своем происхождении из испаноязычной страны. Данная структура представляет собой компромисс между закрытыми и открытыми вопросами в том смысле, что это вопрос открытого типа, имеющий закрытый формат.

В службах NSO могут тестировать различные форматы вопроса, чтобы найти тот, который лучше всего справляется с задачей своевременного сбора необходимых качественных данных без превышения бюджета. В случае правильной реализации использование вопросов нестандартной структуры нередко себя оправдывает.

Письменное поле «Имя»

Вопрос об имени уникален тем, что его можно задать только в открытом формате и тем, что он не предназначен для получения информации, необходимой для целей переписи или опроса. В основном он используется в качестве идентификатора. Имена нередко служат для идентификации респондентов домохозяйства в ходе сбора данных, а также для сопоставления в рамках опросов после переписи (PES, в соответствии с английским акронимом). Имена также используются для нужд индексации в средствах извлечения данных из архивов переписи (см. документ STIC) «Архивация и сохранение данных переписи»).

Еще одна особенность письменных ответов на вопрос об имени в рамках переписи населения состоит в отсутствии необходимости их кодировать. Однако для целей надлежащего использования имен в качестве идентификаторов необходимо устранять ошибки написания, появившиеся на этапе сбора или фиксации данных, поскольку имена часто используются для сопоставления в ходе опроса PES. Сотрудники могут делать ошибки в написании имени, например, во время очных или телефонных опросов. Часто это происходит с именами, имеющими несколько вариантов написания, или с фонетически схожими именами с разным написанием. Например, имя Шон может иметь написание Sean, Shaun, Shawn и Shon, а фамилии Meier, Meyer, Maier, Mayer, Mair и Mayr пишутся по-разному, но произносятся похоже.

После того, как записи переписи переведены в машиночитаемую форму, имена необходимо проиндексировать для обеспечения возможности связывания и извлечения персональных записей. В поле 1 приведено описание Soundex, одной из старейших и наиболее популярных систем фонетической индексации.

Поле 1.

Фонетическая индексация и система Soundex

Система Soundex, разработанная Russell и Odell в 1918 году, считается первой системой фонетической индексации фамилий. К 1930-м годам Soundex уже применялась для индексации данных переписи населения в Соединенных Штатах. Национальное управление архивов и документации США (NARA, в соответствии с английским акронимом), до сих пор применяет данную систему для поиска записей в исторических архивах переписи населения. Благодаря гибкости Soundex основные принципы системы можно адаптировать для различных языков и случаев применения. Ниже представлен испанский вариант Soundex.

Код исходного варианта Soundex состоит из одной буквы и трех цифр. В качестве буквы используется первая буква фамилии, а три цифры выбираются в соответствии с таблицей ниже (если фамилия (last name) недостаточно длинная, для получения трех цифр добавляются конечные нули). В данном алгоритме гласные игнорируются, а согласные группируются со схожими согласными в соответствии с фонетическим местом артикуляции. Дополнительные правила могут применяться для имен с двойными буквами и соседними буквами, имеющими одинаковый номер в Soundex, а также для имен, состоящих из нескольких слов и пр. Обратите внимание, что упоминавшиеся выше фамилии Meier, Meyer, Maier, Mayer, Mair и Mayr имеют один и тот же код Soundex, M600.

ОРИГИНАЛЬНАЯ (АНГЛИЙСКАЯ) СИСТЕМА КОДИРОВАНИЯ SOUNDEX

Код	Буква
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R
Отбрасывать	**A, E, H, I, O, U, W, Y**

Источник: Russell, 1918.

В таблице ниже представлен алгоритм PhoneticSpanish, созданный путем адаптации Soundex для испанского языка. В данной таблице используются цифры 0–9 вместо 1–6; буквы Y и H не отбрасываются; добавлены испанские символы LL, Ñ и RR. Для некоторых сочетаний символов применяются особые правила. Код PhoneticSpanish состоит только из чисел и не имеет ограничения в четыре знака.

ФОНЕТИЧЕСКАЯ КОДИРОВКА ИСПАНСКОГО ЯЗЫКА

Код	Буква
0	P
1	B, V
2	F, H
3	D, T
4	C, S, X, Z
5	L, LL, Y
6	M, N, Ñ
7	K, Q
8	G, J
9	R, RR
Отбрасывать	**A, E, I, O, U, W**

Источник: Amón, Moreno & Echeverri, 2012.

СИСТЕМЫ КОДИРОВАНИЯ

После оцифровки письменных ответов их необходимо классифицировать в соответствии с заранее заданными пронумерованными классами, то есть закодировать. Письменные ответы, полученные с помощью цифровых средств, тоже требуют кодирования. Существует три основных способа кодирования: ручное, кодирование с помощью компьютера и автоматизированное.

Ручное кодирование

Ручное кодирование — наименее сложная система кодирования. Сотрудник, выполняющий кодирование вручную, пользуется при этом только руководством и списком кодов. Такое кодирование могут выполнять сотрудники в процессе сбора данных на местах или специалисты, прошедшие соответствующее обучение, во время обработки данных. В некоторых случаях ручное кодирование выполняется самими респондентами — для этого к анкете может прилагаться список кодов.

С учетом длины письменных ответов их ручное кодирование является нецелесообразным. Ручные системы кодирования нередко являются дорогостоящими и отнимают много времени, даже в небольших странах.

Кодирование с помощью компьютера

Письменные ответы также можно кодировать, используя систему кодирования с помощью компьютера (CAC, в соответствии с английским акронимом). CAC предусматривает интерактивную работу с компьютером для присвоения кодов. Система CAC похожа на ручное кодирование. Однако CAC предоставляет кодировщику доступ к ряду ресурсов на компьютере, например, к экранам подсказки, таблицам принятия решений, вспомогательной информации и др.

Системы CAC более прогрессивны по сравнению с ручными и нередко приносят службе NSO экономию времени и средств. Системы CAC также играют важную роль в оценке и доработке автоматизированных систем кодирования (см. пояснения ниже).

Автоматизированное кодирование

Автоматизированные системы кодирования наиболее развиты в техническом отношении среди всех представленных видов систем. Автоматизированная система предполагает выполнение программы, классифицирующей (т.е. кодирующей) ответы с использованием тщательно проработанных правил. Ответы, которые системе не удалось закодировать, обрабатываются специалистами-кодировщиками с помощью CAC. Ответы, обработанные с помощью CAC, затем итеративно применяются для улучшения способности системы автоматически классифицировать ответы.

Автоматизированные кодировщики уменьшают затраты на обработку и обеспечивают упорядоченное применение правил кодирования. Однако даже самые современные автоматизированные системы требуют некоторого объема ручного кодирования. Ручное кодирование и системы CAC необходимы на начальной стадии разработки автоматизированного кодировщика для обеспечения его корректной работы, а впоследствии — для поддержания в актуальном состоянии.

СОВРЕМЕННЫЕ СИСТЕМЫ КОДИРОВАНИЯ

В рамках современных переписей населения в целях экономически эффективного получения качественных данных используют как автоматизированное, так и ручное кодирование.

Автоматизированная система кодирования состоит из пяти основных частей, или этапов (см. рисунок 2).

Этап 1. Составление словаря

Словари для автоматических систем кодирования подобны спискам кодов, но в отличие от них содержат максимально возможное количество вариантов каждого кода. Словарь для автоматической системы кодирования переписи населения может содержать сотни тысяч вхождений, в зависимости от размера страны. В словаре должны учитываться типичные варианты написания и языковые особенности, в том числе варианты, появившиеся в результате описок и ошибок сканирования из-за неразборчивого или размытого рукописного текста. Например, словарь для условного вопроса о происхождении может включать такие вхождения, как Affrican, Afrcn, Africaine, Africam, Africano, Africsn, Afrika, i-african, Sfrican и т. д. Перечисленные вхождения включают опечатки, ошибочное написание и иностранные слова, эквивалентные слову «африканское», и все они должны быть закодированы одинаково. Словари создаются на протяжении длительного времени — вначале они редко содержат достаточное количество возможных вариантов ответа.

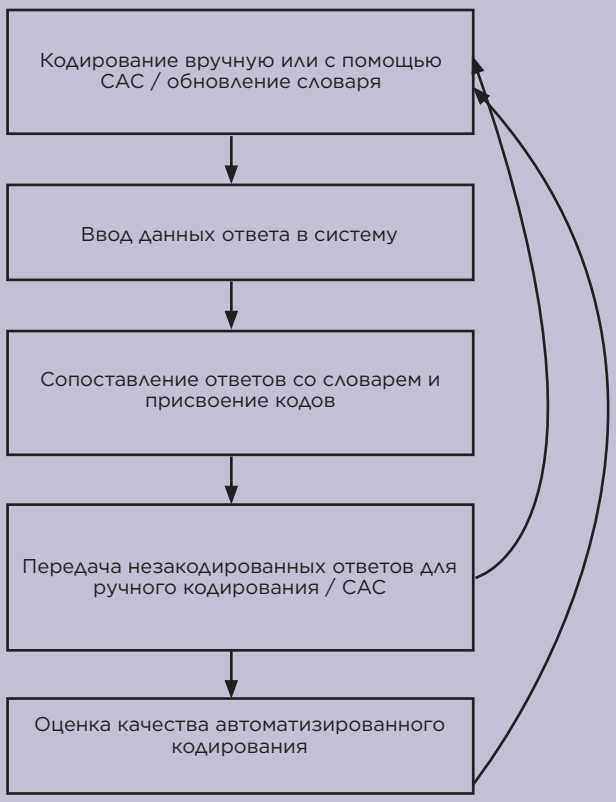
Этап 2. Ввод данных ответа

Оцифрованные данные письменного ответа вводятся в систему без исправлений и редактирования. На этом этапе в письменные ответы разрешено вносить лишь изменения, связанные с удалением недопустимых терминов и символов.

Недопустимые термины включают слова или группы слов, не относящиеся к числу допустимых ответов, и не меняющие смысл ответа после удаления. Например, если на условный вопрос об этнической принадлежности дан ответ «наполовину африканец и наполовину европеец», его необходимо закодировать раздельно как ответ «африканец» и «европеец». Слово «наполовину» и союз «и» можно удалить без изменения смысла подразумеваемого ответа. Однако,

Рисунок 2.

Автоматизированная система кодирования



Источник: U.S. Census Bureau.

если получен ответ «не африканец, европеец», то хотя частица «не» не является допустимым ответом об этнической принадлежности, удалить ее без изменения смысла подразумеваемого ответа нельзя. Термин «не» нельзя игнорировать: его нужно использовать для программирования правила редактирования, согласно которому слово, идущее после «не», тоже должно игнорироваться. Таким образом, в данном примере в качестве допустимого ответа должно быть закодировано только слово «европеец».

К недопустимым символам относятся запятая, апостроф, знак вопроса и восклицательный знак, математические символы и все другие символы, признанные службой NSO в качестве недопустимых. Удаление недопустимых символов применяется только к ответам, извлеченным из оцифрованных бумажных анкет, поскольку в электронных средствах сбора данных, таких как веб-приложения и приложения для планшетов, должны быть разрешены только допустимые символы. Нередко удаление недопустимых символов выполняется при сканировании или оцифровке.

Этап 3. Сопоставление ответов

После удаления недопустимых терминов и символов из записей ответа автоматизированные системы кодирования пытаются сопоставить получившиеся строки текста с кодами службы NSO, используя словарь. В идеале у всех ответов должно быть соответствие в словаре. но на практике это невозможно. На этапе сопоставления ответам, имеющим соответствия в словаре, присваивается код, а ответы без соответствий отмечаются в качестве остаточных.

Этап 4. Сопоставление остаточных ответов

Остаточные ответы переносятся в систему CAC для обработки прошедшими обучение кодировщиками, которые, как правило, специализируются на одном или нескольких конкретных вопросах. Следует иметь ввиду, что остаточные ответы обычно труднее кодировать, так как они более нетипичны в сравнении со среднестатистическими. Поэтому для самых сложных случаев обычно предусматривают особую процедуру — кодировщик обращается за помощью в кодировании или оценке к специалистам в предметной области. После кодирования остаточных ответов в словарь вносят новые соответствия между ответами и кодами службы NSO. Благодаря этому в дальнейшем система сможет автоматически выполнять сопоставление и кодирование таких ответов.

Этап 5. Оценка качества

На данном этапе служба NSO оценивает качество автоматизированной системы с целью ее дальнейшего улучшения. Для оценки качества делается выборка ответов, код которым был присвоен системой автоматически, и эти ответы кодируются вручную наиболее опытными кодировщиками. Затем программисты могут добавить новые правила кодирования, внести корректировки в автоматизированную систему или исправить словарь, если необходимо.

В поле 2 приводится краткое описание процедуры Бюро переписи населения США по кодированию письменных ответов Пробной переписи населения 2019 года на вопросы о происхождении из испаноязычной страны и расовой принадлежности.

Поле 2.

Процедура кодирования и сопоставления Бюро переписи населения США

Бюро переписи населения разработало несколько автоматизированных систем кодирования для различных переписей и опросов. Все эти системы основаны на одних и тех же базовых принципах. Ниже кратко описана процедура кодирования ответов на вопросы Пробной переписи 2019 года о происхождении из испаноязычной страны и расовой принадлежности.

- В рамках Пробной переписи населения 2019 процесс кодирования состоял из двух основных этапов: 1) автоматизированное кодирование и 2) кодирование остаточных ответов.
- В ходе автоматизированного кодирования только что собранные письменные ответы кодировались путем сравнения их с записями в словаре, или «мастер-файле», который содержит сотни тысяч ранее закодированных письменных ответов, накопившихся за несколько переписей и опросов, с соответствующими кодами. При наличии соответствия выполнялось присвоение кода. Письменным ответам без соответствия должен был вручную присвоить код кодировщик, прошедший необходимое обучение.
- Процесс остаточного кодирования включал три основных вида деятельности:
  a. Производственное кодирование — ручное кодирование остаточных ответов.
  b. Проверочное кодирование — другой специалист выполнял повторное кодирование выборки случаев.
  c. Оценка — расхождения между производственным и проверочным кодированием устранялись третьим кодировщиком более высокой квалификации (оценщиком).
- После того, как остаточные ответы прошли кодирование и контроль качества, их внесли в мастер-файл для улучшения процесса автоматизированного кодирования путем создания новых возможных соответствий для будущих письменных ответов.

ЗАКЛЮЧЕНИЕ

Разработка эффективных систем кодирования письменных ответов может стать одной из самых обременительных и сложных задач в рамках переписи населения. Однако при надлежащем планировании и эффективном использовании технологий объем работы службы NSO и затраты можно уменьшить. Правильно реализованная автоматизированная система кодирования с элементами ручного кодирования позволяет максимально автоматизировать процесс без принесения в жертву качества.

ЛИТЕРАТУРА

Amón, I., Moreno, F., and Echeverri, J., "Algoritmo fonético para detección de cadenas de texto duplicadas en el idioma español," *Revista Ingenierías, Universidad de Medellín*, 11(20), 127–138, 2012.

Bethlehem, J., "Applied Survey Methods: A Statistical Perspective," Wiley series in survey methodology, Hoboken, N.J., Wiley, 2009.

Campanelli, P., Thomson, K., Moon, N., and Staples, T., "The Quality of Occupational Coding in the United Kingdom," In Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., and Trewin, D., eds., *Survey measurement and process quality,* pp. 437–453, New York, Wiley, 1997.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., and Tourangeau, R., Survey methodology 2nd ed., Wiley series in survey methodology, Hoboken, N.J., Wiley, 2009.

Lyberg, L., and Kasprzyk, D., "Some aspects of post-survey processing," In Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., and Trewin, D. eds., *Survey measurement and process quality,* pp. 353–370), New York, Wiley, 1997.

Rea, L., and Parker, R., "Designing and Conducting Survey Research: A Comprehensive Guide," fourth ed., San Francisco, CA, Jossey-Bass, a Wiley brand, 2014.

Russell, U.S. Patent No. 1261167, 1918.

Saris, W., and Gallhofer, I., "Design, Evaluation, and Analysis of Questionnaires for Survey Research," second ed., Wiley series in survey methodology, Hoboken, New Jersey, Wiley, 2014.

U.S. Census Bureau, American Community Survey, Design and Methodology, Washington, D.C., U.S. Department of Commerce, Economics and Statistics Administration, U.S. Census Bureau, 2009.

Wolf, C., Joye, D., Smith, T., and Fu, Y., eds., "The Sage Handbook of Survey Methodology," Los Angeles, Sage Publications, 2016.

Серия «Избранные темы международных переписей населения» (STIC) публикуется в рамках Международных программ отделения по народонаселению Бюро переписи населения США. Агентство США по международному развитию финансирует подготовку документов серии STIC и двустороннюю поддержку статистических организаций, которые предоставляют информацию для авторов. Фонд Организации Объединенных Наций в области народонаселения участвует в подготовке содержания и обнародовании документов STIC, способствуя их распространению среди более широкой аудитории.