

**КОМИССИЯ ПО СТАТИСТИКЕ и  
ЭКОНОМИЧЕСКАЯ КОМИССИЯ ДЛЯ ЕВРОПЫ**

**КОМИССИЯ  
ЕВРОПЕЙСКИХ СООБЩЕСТВ**

**КОНФЕРЕНЦИЯ СТАТИСТИКОВ ЕВРОПЫ**

**ЕВРОСТАТ**

**Совместный рабочий семинар ЭКЕ и ЕВРОСТАТа  
по конфиденциальности статистической информации**  
(Скопье, бывшая югославская республика Македония,  
14-16 марта 2001 г.)

Рабочий доклад №21

Тема II: Влияние новых технологических разработок в программном обеспечении, средствах связи и вычислительных процессах на SDC (Контроль за соблюдением конфиденциальности статистической информации)

**СИСТЕМА «АМЕРИКЭН ФЭКТ ФАЙНДЕР»: ДЕЯТЕЛЬНОСТЬ БЮРО ПЕРЕПИСИ США  
ПО УДОВЛЕТВОРЕНИЮ ЗАПРОСОВ ПОЛЬЗОВАТЕЛЕЙ И ЗАЩИТЕ  
КОНФИДЕНЦИАЛЬНЫХ СТАТИСТИЧЕСКИХ ДАННЫХ**

**Представленная работа**

Представлена Бюро переписи США<sup>1</sup>

**Аннотация:** «Америкэн Фэкт Файндер» (American FactFinder или AFF) – это новая система Бюро переписи Соединенных Штатов по распространению данных в интерактивном режиме, обеспечивающая доступ к данным переписи 1990 г. десятилетней периодичности, генеральной репетиции переписи 2000 г., обследования американских общин, экономической переписи 1997 г. и, со временем, переписи 2000 г. Большая часть общедоступных файлов данных представляет собой итоговые файлы с матрицами агрегированных данных. Система способна вырабатывать табулярные данные на основе запрошенных файлов микроданных, находящихся за сетевым экраном (3-й уровень). Распространение табулярных данных в интерактивном режиме на основе запросов файлов полных микроданных требует специальных способов ограничения риска нарушения конфиденциальности. Принципы и методы, описываемые в данной работе, применяются для защиты файлов полных микроданных переписи 2000 г. третьего уровня доступа.

**I. ВВЕДЕНИЕ**

1. Бюро переписи США является важнейшим сборщиком и распространителем своевременных, значимых и точных данных о населении и экономике Соединенных Штатов. Проведя более 100 ежегодных обследований и 20 переписей десятилетней периодичности, начиная с первой переписи 1790 г., Бюро переписи распространяет официальную информацию по населению, компаниям, организациям и отраслям экономики США. Бюро переписи гарантирует конфиденциальность данных респондентов в течение 72 лет, как того требует федеральный закон (Статья 13, Раздел 9 Кодекса США). Помощь со стороны граждан, предприятий и других респондентов, предоставляющих соответствующие сведения, необходимые для разработки статистических данных, во многом зависит от того, насколько успешно Бюро переписи находит оптимальное равновесие между необходимостью защищать права отдельных лиц на конфиденциальность частных сведений и необходимостью распространять своевременную и полезную информацию. В целях обеспечения конфиденциальности сведений физических и юридических лиц при предоставлении данных внешним по отношению к Бюро переписи пользователям, Бюро переписи разработало свод правил и методов, ограничивающих

---

<sup>1</sup> Подготовил Сэм Хоуала (e-mail: sam.hawala@census.gov).

риск нарушения конфиденциальности. Целью данной работы является описание этих принципов и методов, используемых в системе «Америкэн Фэкт Файндер».

## II. СИСТЕМА «АМЕРИКЭН ФЭКТ ФАЙНДЕР»

2. До 1960 г. Бюро переписи выпускало данные только в виде таблиц. Табулирование происходило на блоковом уровне собранных данных с помощью кратких статистических форм данных стопроцентно десятилетней периодичности. Эти данные по выборкам десятилетней периодичности объединялись на трактовом уровне в длинные статистические формы. По мере внедрения недорогих и мощных компьютерных систем и средств хранения данных обнаружилось, что некоторые исследовательские задачи лучше всего решать исключительно на базе микроданных. В 1963 г. на основе выборки из переписи 1960 г. десятилетней периодичности был разработан и распространен первый общедоступный файл выборочных микроданных. Образцы микроданных общего доступа содержат ответы респондентов по анкетам, из которых в целях защиты конфиденциальности респондентов изъяты уникальные идентификаторы (имена, адреса и т.д.).

3. В рамках инициативы администрации Клинтон, направленной на повышение эффективности и общедоступности правительственных органов, Бюро переписи объявило в октябре 1998 г. о введении новой системы распространения данных через Интернет, которая значительно расширяет возможности доступа пользователей к огромным информационным базам агентства. Эта новая система является дополнением к существующему интернетовскому сайту Бюро переписи, впервые предоставляя общественности доступ к крупнейшим программам и базам данных Бюро переписи в интерактивном режиме. Эта новая система получила название American FactFinder (AFF). Она создается по контракту с Бюро переписи корпорацией “IBM Global Services Corp.” (“Ай-Би-Эм Глобал Сервисиз Корпорейшн”), генеральным подрядчиком, отвечающим за общую компоновку системы и разработку пользовательского интерфейса.

4. Первые данные, распространенные через систему AFF, представляли собой предварительные результаты экономической переписи 1997 г., переписи населения 1990 г. с файлами данных по жилищным условиям, пробные и демонстрационные данные обследования американских общин, а также результаты генеральной репетиции переписи 2000 г., проводившейся в 1998 г.

5. Полные разработанные данные переписи 2000 г. станут доступны через систему AFF в январе 2001 г., начиная с суммарных показателей по миграции населения страны и подробных суммарных показателей (вплоть до блокового уровня) по миграции из одних округов в другие. Блоки переписи представляют собой самые мелкие территориальные единицы, по которым Бюро собирает и табулирует данные переписи десятилетней периодичности. Границы блоков определяются реальными физическими границами (улицы, дороги, железнодорожные магистрали, а также источники воды) и/или культурными признаками.

6. При определенных, указанных в данной работе ограничениях, пользователи системы AFF могут задавать свои собственные географические районы для нестандартного табулирования. Нестандартное табулирование и табулирование для мелких географических районов вводят повышенный риск раскрытия индивидуальных данных. Угроза нарушения конфиденциальности частной информации может повлиять на готовность населения к сотрудничеству при переписях и обследованиях, проводящихся Бюро переписи. Поэтому Бюро переписи в течение последних лет прикладывает значительные усилия к защите конфиденциальных данных. Целью поиска и разработки описываемых в данной работе правил и методов ограничения риска нарушения конфиденциальности является сохранение прочной репутации Бюро переписи в отношении обеспечения конфиденциальности.

### **III. ПРАВИЛА И МЕТОДЫ ОГРАНИЧЕНИЯ РИСКА НАРУШЕНИЯ КОНФИДЕНЦИАЛЬНОСТИ**

7. Посредством AFF общественность сможет получать сводные, введенные в систему данные переписи 2000 г. через Интернет. Данные будут распространяться в виде таблиц на экране монитора пользователя и в виде эквивалентного таблице “мягкого” (копируемого) файла. После того, как Бюро переписи определило таблицы и разрешило пользователям допуск к своду соответствующих таблиц, такой результирующий свод готовых таблиц называется “данными 2-го уровня”. Когда внешние пользователи задают свои собственные таблицы, результирующие таблицы называются “данными 3-го уровня”. Лежащие в основе таблиц как 2-го, так и 3-го уровня данные представляют собой “микроданные” об отдельных лицах и домашних хозяйствах. Для обеспечения конфиденциальности микроданных при доступе внешних пользователей к табулярным данным Бюро переписи применяет перекодировку данных, перестановку данных и фильтрацию запросов на данные как 2-го, так и 3-го уровня. Все таблицы 2-го уровня одобрены Советом Бюро переписи по надзору за конфиденциальностью. Сводные данные 3-го уровня предоставляются только при соблюдении правил, ограничивающих риск нарушения конфиденциальности.

#### **III.1 Перекодировка данных**

8. Такие переменные как национальность, род занятий, отрасль экономики, латино-американское происхождение, район компактного проживания перекодируются в новые, менее подробные переменные. Все такие непрерывные переменные как доход домашнего хозяйства/семьи, виды индивидуальных доходов, затраты на электроэнергию, газ, воду, топливо, налог на собственность, выплаты по закладным, общая арендная плата, подвергаются перекодировке по верхней границе. Отбрасываемые значения при перекодировке по верхней (иногда по нижней) границе определяются структурой распределения подлежащей перекодировке переменной.

9. Перекодированные переменные вводятся в файлы системы AFF и становятся доступны в зависимости от характера запроса и самого запрашивающего. Внутренние пользователи имеют доступ ко всем переменным и записям, необходимым им в работе. Внешние пользователи отсылаются к перекодированным переменным в зависимости от таких затрагивающих конфиденциальность вопросов как:

- группы населения, географический район, запрошенные переменные и
- соблюдение правил и пороговой численности населения, определяемых Бюро переписи для таблиц в целом и табличных ячеек, которые должны удовлетворять критериям конфиденциальности.

#### **III.2 Перестановка данных**

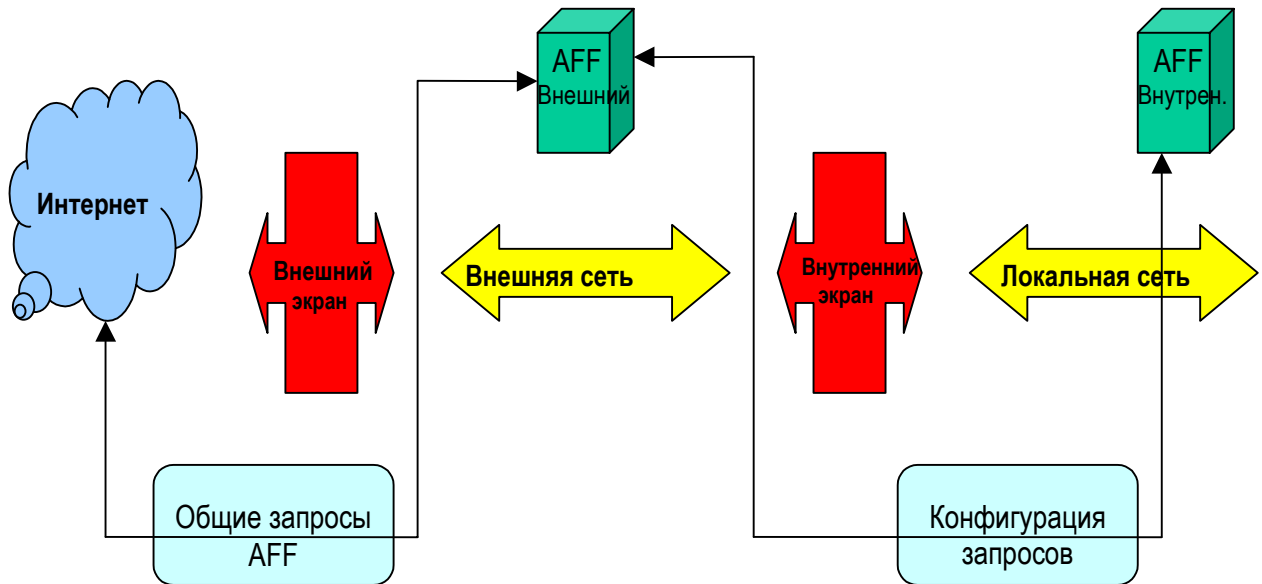
10. Метод перестановки применялся к данным переписи населения 1990 г. и данным по жилищному фонду. Этот метод будет применяться к данным – на сто процентов по коротким формам – переписи 2000 г. и, независимо от этого, к выборочным данным – по длинной форме – переписи 2000 г. Метод заключается в перестановке пар записей, отобранных как представляющие самый высокий риск нарушения конфиденциальности. В частности, для перестановки отбираются записи, уникальные с точки зрения комбинации переменных. Между двумя различными территориальными единицами переставляются все записи по домашним хозяйствам.

11. Переменные, делающие запись уникальной, называются ключевыми переменными. Для перестановки отбирается запись с вероятностью, обратно пропорциональной размеру блока. Записи по домашним хозяйствам с уникальными расовыми признаками (категориями) в блоке имеют большую вероятность быть переставленными. Переставленные записи совпадают по ряду демографических характеристик, но находятся в разных блоках переписи.

12. В системе AFF для распространения будут использоваться файлы с переставленными данными. Все таблицы, подлежащие широкому распространению (в печатном варианте, на магнитных носителях, через систему AFF и т.д.) вне Бюро переписи, будут разрабатываться из файлов с переставленными данными.

### III.3 Технология AFF

13. Интернетовский сетевой экран Бюро переписи сконфигурирован на допуск в интерактивном режиме только тех запросов, которые поступают от единственного внешнего сервера AFF на внутренний сервер AFF. Любые запросы на внутренний сервер, поступающие от каких-либо внешних устройств, блокируются сетевым экраном.



14. Все запросы от внешних пользователей направляются через канал «Общие запросы AFF», как показано на рисунке. По определению этот маршрут не может использоваться для передачи конфиденциальной информации. Внешний сервер принимает запросы на данные 3-го уровня и пересылает их по маршруту «Конфигурация запросов». Этот маршрут также не используется для конфиденциальной информации. Сетевые пакеты на этом втором маршруте не могут быть отслежены внешним пользователем, поскольку маршрутизатор Интернета не пропускает эти пакеты в Интернет. Ни один сетевой протокол Интернета не может связаться с внутренним сервером, так как сетевой экран разрешает связь только между двумя серверами системы AFF. Нет необходимости и кодировать информацию, идущую по этим двум маршрутам, так как в ней нет ничего конфиденциального.

15. Большая часть файлов данных, доступных через AFF, являются сводными файлами с матрицами агрегированных показателей, которые находятся на 1-ом и 2-ом уровнях. Отличительной особенностью системы AFF является ее способность создавать задаваемые пользователем таблицы на основе находящихся за сетевым экраном файлов микроданных. Это 3-й уровень. Через 3-й уровень пользователь может получить очень подробные данные по детализированным географическим районам. Далее мы обсудим правила ограничения риска нарушения конфиденциальности, предназначенные для запросов на 3-й уровень. Эти правила действуют как ряд фильтров. Имеется два вида фильтров: фильтр запросов и фильтр результатов.

#### III.3.1 Фильтры запросов

16. Задача фильтров запросов – обнаружение таких запросов, которые не могут до их передачи на исполнение пройти ограничители. Этот фильтр экономит ресурсы системы 3-го уровня и время

внешних пользователей, относительно оперативно сообщая им, удовлетворяют ли их запросы критериям ограничения риска нарушения конфиденциальности.

- ◆ Перекрещивающиеся таблицы должны строиться по географическим районам и/или заданному Бюро переписи списку негеографических переменных.
- ◆ Географические переменные в запросах должны удовлетворять минимальным пороговым показателям. Система определяет, запрашиваются ли мелкие районы –блоки, группы блоков или заданные пользователем территориальные единицы с численностью населения меньше среднего (4060 в 1990 г.), средние по численности населения районы (от 4060 до 99 999 жителей) или крупные районы (население 100 000 и выше).
- ◆ В зависимости от численности населения в районе или запрошенных районов система разрешает использование соответствующих комбинаций коротких, средних или длинных рядов заданных категорий национальности, латино-американского происхождения, мест компактного проживания и других переменных для перекрестного табулирования. Доступ возможен только к перекодированным по верхней границе переменным.
- ◆ При предоставлении значений ячеек для раздробленного района раздробленность игнорируется и сообщаются общие данные по району.
- ◆ Максимальное количество переменных, используемых для построения общей таблицы, равно трем, исключая географическую переменную.
- ◆ В зависимости от переменных перекрестного табулирования пользователь должен выбрать из списка допустимые параметры значений для ячейки (среднее арифметическое, медианы, ...).
- ◆ Параметры предоставляются только при наличии соответствующих рассчитанных показателей.
- ◆ Если расчетное время исполнения или объем выходных данных превышает порог, установленный Бюро, запрос отклоняется.

17. Если фильтр запроса отклоняет первоначальный запрос пользователя, пользователь имеет возможность изменить пороговый показатель численности населения или выйти из сети. Например, если пользователь посылает запрос на информацию по национальной принадлежности, который удовлетворяет пороговому значению средней численности населения в данном районе, а пороговое значение численности населения района оказывается трактом, тогда пользователь имеет возможность воспользоваться соответствующей перекодировкой без учета национальной принадлежности, подходящей для данного порогового значения численности населения в небольшом районе.

18. Если запрос удовлетворяет всем ограничительным правилам фильтра запросов, запрос пересылается из внешнего сервера на внутренний, находящийся за сетевым экраном сервер и к файлам полных микроданных для процесса вычисления. Файлы полных микроданных содержат все уже заданные категории национальной принадлежности, латино-американского происхождения, мест компактного проживания, а также переменные модифицированных данных выборки.

### **III.3.2 Фильтр результатов**

19. Фильтр результатов обеспечивает последнюю проверку показателей ячеек полученной таблицы. Таблица предоставляется внешнему пользователю, если и только если она удовлетворяет всем критериям ограничения риска нарушения конфиденциальности, определенным для фильтра результатов.

- ◆ При запросе территориальных подтаблиц, критерии ограничения риска нарушения конфиденциальности применяются отдельно к каждой территориальной единице соответствующей подтаблицы, как если бы пользователь прислал отдельные запросы по каждой территориальной единице.
- ◆ Значение медианы в ячейке запрошенной таблицы не может быть меньше параметра, установленного Бюро переписи.
- ◆ Среднее значение в ячейке запрошенной таблицы не может быть меньше параметра, установленного Бюро переписи.

- ◆ Отношение количества ячеек с единичным подсчетом к общему количеству ячеек в запрошенной таблице не может превышать параметр, установленный Бюро переписи.
- ◆ Если пользователь задает при запросе перекодировку, система 3-го уровня прежде всего рассчитывает результаты без учета требования по перекодировке и сопоставляет их с критериями по ограничению риска. Затем, если результаты удовлетворяют этим критериям, они агрегируются согласно затребованной перекодировке.

20. Бюро переписи продолжит опробирование вышеописанных фильтров для третьего уровня системы AFF на данных переписи 2000 г. десятилетней периодичности как только эти данные станут доступны для соответствующей корректировки.

### **Литература**

Заяц, Лаура, и Роуланд, Сандра (2000), «Ограничение риска нарушения конфиденциальности при переписи 2000 года», Материалы сектора по исследованию методики обследований, Американская статистическая ассоциация, выходит в свет.