Информационные технологии

УДК 004.652.4:929.52

Построение реляционной модели данных о жителях Российской империи по оцифрованным документам российских переписей

Н. Е. Брилёва, А. С. Панкратов

Кафедра информационных технологий
Российский университет дружбы народов
ул. Миклухо-Маклая, 6, г. Москва, Россия, 117198

Описывается технология построения базы данных, содержащей материалы переписей населения Российской империи, на примере ревизских сказок XVIII–XIX веков. База данных строится на основе оцифровки архивных материалов переписей, сделанной в формате Excel. Описывается алгоритм распознавания смысловых конструкций записей в переписных документах, выделения атрибутов и правила заполнения реляционных таблиц. Алгоритм приводится для конкретных переписей (4 и 5 ревизий), однако подобные алгоритмы могут быть разработаны и для других ревизий, что позволит выстроить непрерывную цепочку исторических переписных ведомостей, переведенную в единый формат базы данных. Наличие такой цепочки может оказать существенную помощь в генеалогических исследованиях.

Ключевые слова: база данных, полуструктурируемые данные, оцифровка, генеалогия, перепись, ревизия, ревизская сказка.

1. Введение

С ростом количества самых разнообразных данных возрастает потребность в организации информации таким образом, чтобы можно было получать достоверные ответы на интересующие пользователя вопросы и эффективно управлять данными.

Уже много лет эти задачи решаются при помощи информационных систем. Данные, с которыми работают информационные системы, могут значительно различаться по степени структурированности. С одной стороны, данные, хранящиеся в традиционных реляционных базах данных, имеют строгую и определенную структуру. С другой стороны, можно назвать ряд данных, не обладающих четкой структурой. Таковыми являются, к примеру, электронные книги, фильмы, печатные документы, презентации, рисунки, рентгеновские снимки, отпечатки пальцев, фотографии, видеоролики, архивы камер наблюдения и многое другое. Между этими двумя крайностями находится большой объем так называемых полуструктурируемых (или слабоструктурируемых) данных. К полуструктурируемым относят такие данные, в которых можно выделить некоторую структуру (в отличие от аудио- или видео-данных), но структура этих данных не достаточно строгая для хранения их в традиционных системах (реляционных, объектно-ориентированных), либо эта структура способна динамически изменяться с течением времени. Примерами полуструктурируемых данных могут быть форматированные тексты, HTML-страницы, документы со структурированными фрагментами. В настоящее время имеется ряд теоретических исследований и прикладных разработок, позволяющих оперировать с такого рода данными, в частности, извлекать из них нужную информацию [1–3].

Настоящая статья посвящена полуструктурируемым данным специального вида, а именно, данным, полученным в результате оцифровки материалов переписей населения Российской империи, взятых из архива. Эти материалы играют важнейшую роль при проведении генеалогических исследований, восстановлении

Статья поступила в редакцию 27 декабря 2011 г.

родословных линий. Подобные исследования сейчас становятся всё более популярными, многие хотят восстановить историю своей семьи, своего рода. Большинство государственных архивов в настоящее время открыто для широкого круга исследователей (как профессионалов, так и любителей), однако людям, которые хотят узнать больше о своих предках, приходится прикладывать огромные усилия для того, чтобы найти какие-то сведения в различных архивных документах. Это связано с массой факторов, в частности, с географической удаленностью региональных архивов. Кроме того, в связи с рядом исторических факторов колоссальное количество архивных документов оказалось безвозвратно утраченным. Но если исследователь-любитель и получит доступ к документам, в которых имеется информация о его предках (например, к материалам переписей), он может потратить колоссальный объем времени на то, чтобы просто прочитать записи в переписях и разобрать почерк того приказчика или старосты, который проводил перепись.

Но наиболее сложным является процесс идентификации выбранного человека и его сородичей в других, более ранних (поздних) переписях. Если даже исследователь изучит структуру каждой конкретной переписи, аккуратно проследит записи о каждом конкретном предке в этих переписях и ничего не перепутает и не ошибется (что требует большой концентрации и аккуратности в работе), то ему потребуются долгие дни, а может и недели, чтобы продвинуться в своих изысканиях.

Уже давно перед разработчиками и программистами поставлена задача автоматизировать процесс поиска в архивах информации о человеке. В частности, разработана и успешно функционирует система поиска сведений о погибших и пропавших без вести в годы Великой Отечественной войны — на основе данных Центрального архива Министерства Обороны [4].

Однако относительно данных переписей населения Российской империи в контексте построения родословных связей задача автоматизации поиска не является тривиальной. Основная сложность здесь связана с необходимостью структуризации информации для дальнейшей обработки тех данных, которые представлены в архивах.

Среди множества архивных документов, используемых в генеалогических исследованиях, можно выделить следующие основные категории [5]:

1) писцовые книги (поземельные переписи) XV–XVII веков;
2) подворные переписи 1646–1717 гг.;
3) ревизские сказки (подушные переписи) 1719–1858 гг.;
4) церковные исповедные росписи;
5) церковные метрические книги, включавшие метрические тетради рождения, брака, смерти;
6) материалы первой всеобщей переписи населения Российской империи 1897 г.

Данные всех вышеназванных архивных документов по своему характеру являются полуструктурируемыми, причём, как правило, в более поздних по времени документах структура проявляется более отчётливо, чем в ранних.

Документы переписей помимо списков жителей населённых пунктов фиксировали также и родственные связи в пределах каждой семьи. Эта информация представляется чрезвычайно важной для генеалогических исследований, поскольку она позволяет строить восходящие цепочки от сына к отцу, деду и т.д. Однако при этом возникает задача идентификации конкретного персонажа (и членов его семьи) в иных переписных документах, близких по времени, но, возможно, имеющих уже несколько иную структуру.

Для автоматизации построения родословных линий в качестве первоначальной задачи требуется адекватная оцифровка имеющихся переписных архивных документов и их последующее объединение (интеграция) в единое целое.

В настоящей статье предлагается один из путей решения данной задачи, основанный на построении базы данных, в которой содержатся информация о людях, полученная из формуляров ревизских сказок, содержащих данные о крестьянах, жителях сел и деревень.

Мы ограничились рассмотрением одного села, но возможно расширение географических масштабов путём введения дополнительных отношений и атрибутов.

2. Ревизские сказки и их формуляры

С 1719 по 1858 г. было проведено десять ревизий (переписей), основной целью их проведения был учёт населения, подлежащего налогообложению. Как правило, учету подлежали все категории податного населения (а их насчитывалось более 100 наименований), традиционно объединяемые в более крупные группы крестьян (государственных, дворцовых и частновладельческих), посадских людей и купцов. Во всех без исключения ревизиях не участвовали дворяне, духовенство, отставные солдаты и драгуны, а также состав действующей армии и флота. Единицей переписи была так называемая «ревизская душа», сохранявшая свою актуальность как единица налогообложения вплоть до следующей ревизии, несмотря на возможное изменение социального статуса, забор в рекруты или смерть.

В формуляре переписи населения обычно указывались следующие данные: номер двора, жители двора их возраст и родственные связи, иногда приводились дополнительные характеристики людей: вдова/вдов, солдатка, девка и т.д. У женщин встречалась характеристика «старинная того же села», что означает, что она проживала в этом же селе и до замужества. В 1-й, 2-й и 6-й ревизиях информация о женщинах вообще не указывалась.

Четкого формуляра ревизской сказки долгое время не существовало, его структура могла меняться от переписи к переписи. Так, в 1-й и 2-й ревизиях (1719 и 1747 гг.) данные о жителях дворов записывались сплошным текстом, при этом запись о каждом дворе могла начинаться словами «Во дворе» либо стилизованной буквой «В», взятой в кружок. В 3-й ревизии (1763 г.) данные о жителях (мужчинах) уже организуются в таблицы со следующими четырьмя колонками: в 1-й (основной) колонке записывались сами жители (для хозяина двора — имя, отчество и фамилия (если была), для членов его семьи — только имена с указанием степени родства), во 2-й и 4-й колонках записывался возраст жителя, соответственно, на предыдущую и текущую перепись, в 3-й колонке фиксировались выбывшие из списка с момента предыдущей переписи, с указанием года и причины выбытия. Женщины, впервые начавшие упоминаться именно в этой ревизии, записывались в первую колонку, с указанием степени родства и возраста в этой же колонке. В таблицах 4-й и 5-й переписей (1782 и 1795 гг.) женщины уже упоминаются наряду с мужчинами, только колонки «возраст на предыдущую перепись», «выбытие» и «возраст на текущую перепись» для них организуются отдельные, в правой части таблицы.

Только к 7-й ревизии (1816 г.) структура формуляра сложилась окончательно. С этого момента ревизская сказка содержит в себе следующие элементы: название (год, месяц, число подачи, губерния, уезд, село, владелец крепостного селения, если сказка дана по помещичьему селу), графу с номером семьи, цифры состава семьи «по последней ревизии», число и состав «выбывших» и «прибывших» лиц, временно отсутствовавших, и итоговые данные «ныне налицо». В сказку записывалась одна семья за другой: на левой половине разворота — мужчины, на правой — женщины. Информация о женщинах при этом упрощается: пропадают графы «выбытие» и «возраст на предыдущую перепись». В 7-й и 8-й ревизиях у женщин также не указывались отчества [5].

3. Оцифровка переписных документов

Самый тривиальный способ оцифровки какого-либо документа — это его сканирование, перевод в формат изображения. Но подобный способ не позволяет осуществлять в оцифрованном документе какой-либо автоматический поиск. Для организации автоматического поиска в архивных документах необходим их перевод

(хотя бы частичный) в иной формат — текстовый, табличный, формат базы данных и пр. Технология сканирования с последующим автоматическим распознаванием текста здесь также неприменима, поскольку старинные документы писались от руки, а в документах XVIII века и более ранних использовалась старинная скоропись, с сокращениями, титлами и надстрочными буквами, прочитать которую возможно только при некотором навыке.

В настоящей работе предполагается, что архивные документы оцифровываются вручную. Табличная структура данных ревизий (начиная с 3-й ревизии) позволяет использовать в качестве формата оцифровки электронные таблицы Excel. При этом производится перевод старинной скорописи на современный русский язык, однако с максимальным сохранением исходной орфографии и пунктуации.

На рис. 1 представлен оригинальный лист ревизской сказки 1782 г. (4-й ревизии), табл. 1 представляет его оцифрованный Excel-вариант.

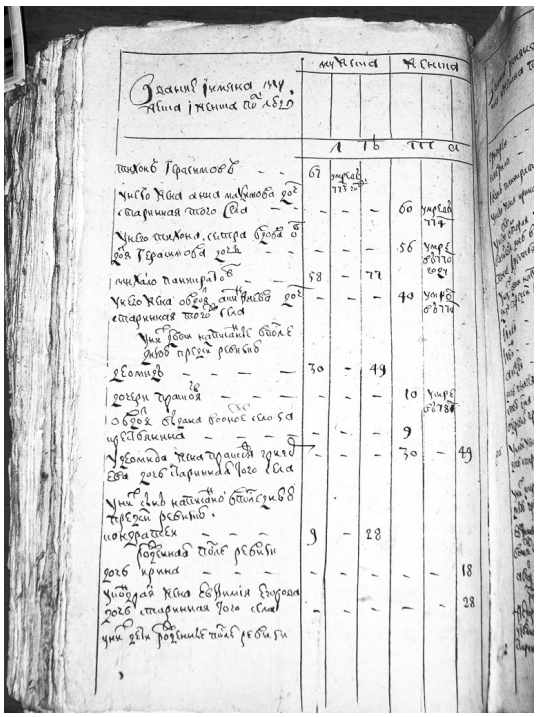

Рис. 1. Формуляр 4-й ревизии (1782 г.)

Как видно из табл. 1, имена-отчества жителей, их родственные связи, а также некоторые дополнительные характеристики (такие, как «вдова», «старинная того же села») попадают в одну графу Excel-таблицы. Для задачи поиска конкретного персонажа и автоматического выявления его родственных связей возникает проблема распознавания в этой графе всех названных характеристик переписанных жителей (в дальнейшем — персонажей) и размещение их в реляционные таблицы базы данных. Если проделать это для каждой ревизии, то мы получим структурированную хронологическую цепочку оцифрованных переписей, организованных в единую базу данных. И эта база данных позволит нам для каждого конкретного персонажа, представленного в переписях, выявить его родословные линии на протяжении примерно полуторавекового периода.

Все 10 ревизий можно разбить на 4 группы сходных по стилю и формуляру: 1-я и 2-я, 3-я и 6-я, 4-я и 5-я, с 7-й по 10-ю. Каждая группа характеризуется своими правилами распознавания характеристик персонажей. В данной работе мы ограничимся рассмотрением 4-й и 5-й ревизий (1782 и 1795 гг.)

Таблица 1

Персонажи	возраст на пред. перепись (муж.)	выбытие (муж.)	возраст на эту перепись (муж.)	возраст на пред. перепись (жен.)	выбытие (жен.)	возраст на эту перепись (жен.)
Тихон Герасимов	67	умре в 773 году				
У него жена Анна Максимова дочь старинная того села				60	умре в 774	
У него Тихона сестра вдова Овдотья Герасимова дочь				56	умре в 770 году	
Михайло Паникратов	58		77			
У него жена Овдотья Акинфиева дочь старинная того же села				44	умре в 770	
У них дети написанные в последней пред сей ревизии Деомид	30		49			
дочери Прасковья				10	умре в 781	
Овдотья выдана в оное же село за крестьянина				9		
У Деомида жена Прасковья Григорьева дочь старинная того села				30		49
У них сын написанный в последней пред сей ревизии Кондратей	9		28			
рожденная после ревизии дочь Ирина						18
У Кондратия жена Евфимия Егорова дочь старинная того села						28

4. Перевод оцифрованных документов в формат реляционной базы данных

С помощью инструментов экспорта/импорта оцифрованный Excel-вариант ревизской сказки (см. табл. 1) преобразуется в промежуточную реляционную таблицу базы данных MS SQL Server. Пустые ячейки при этом заполняются значениями Null.

Следует заметить, что в разных городах и уездах даже в рамках одной ревизии способ описания характеристик персонажей мог отличаться (зависело от конкретного писаря), хотя смысл этих характеристик был одинаковый. В частности, возраст до 2 лет в исходных исторических документах мог прописываться прописью: «года», «полгода», «4 месяцев», «года и 5 месяцев»... Для облегчения задачи распознавания целесообразно стандартизировать некоторые выражения, используемые для описания персонажей и привести их к стандартному виду.

Примеры характеристик, приводимых к стандартному виду:

1) «у него жена» — «его жена»; «жена его»;
2) «у него сестра» — «его сестра»; «сестра его»;
3) «у них дети» — «их дети», «дети их», «ихние дети»;
4) данные о возрасте приводятся к числовому формату.

Приведение к стандартному виду производится встроенными средствами языка T-SQL, с помощью специально определенной функции на основе таблицы словосочетаний-синонимов. Данные перезаписываются во вторую промежуточную таблицу с добавочным первым столбцом (ключевым), содержащим порядковые номера персонажей.

5. Алгоритм распознавания характеристик персонажей

В настоящее время существует ряд алгоритмов и программ автоматической обработки текста, включающих семантический анализ и распознавание смысловых конструкций [6]. Однако записи в ревизских сказках обладают рядом специфических особенностей (например, древнерусские языковые обороты и старинная орфография), и количество используемых в них языковых конструкций невелико. В связи с этим для поставленной задачи выделения характеристик из переписных документов представляется целесообразным разработать специальный алгоритм, задача которого — сформировать на базе ревизских записей реляционные таблицы с характеристиками персонажей.

Информация, присутствующая в сказках 4-й и 5-й ревизий, позволяет выделить следующие характеристики: имя, отчество, фамилия (не всегда), пол, возраст, родственник, степень родства, дополнительные характеристики (вдов, вдова, солдатка и пр.) В связи с возможным многозначным характером дополнительных характеристик они выделяются в отдельную таблицу. Итак, алгоритм будет формировать две реляционные таблицы: «Персонажи» с полями: номер двора, номер персонажа, имя, отчество, фамилия, пол, год рождения, номер родственника, степень родства и «Доп_характеристики» с полями: номер персонажа, характеристика. При этом будет использоваться вспомогательная таблица — словарь имён с указанием соответствующего пола.

Определим для начала основные правила и типы записей, которые будем использовать в алгоритме.

Основные правила:

1) глава семьи: имя в именительном падеже стоит на первом месте, отчество присутствует обязательно;
2) имя персонажа в описании идет в именительном падеже и точно совпадает с одним из имен из словаря имен.

Типы записей выпишем в табл. 2.

Таблица 2

Тип	**Описание**	**Образцы записей**
1	Глава семьи	Тихон Герасимов сын Кузнецов
		Петр Васильев вдов
		Солдатка Прасковья Григорьева Шмелева
2	Жена главы семьи	У него Петра жена Овдотья Акинфиева дочь
		У него вторая жена Ирина Максимова старинная того же села
3	Брат главы семьи	У него брат Григорий Герасимов вдов
4	Сестра главы семьи	У Тихона сестра вдова Анна Герасимова дочь
5	Дети (имя с предварительным текстом, как правило – сыновья)	У них дети, записанные в последней пред сей ревизии Иван
		Рожденный после ревизии Илья
		У него сын Савва
6	Дочери (имя с предварительным текстом)	Дочь Прасковья выдана в замужество в село Подолец за крестьянина
		У них дочери записанные в последней пред сей ревизии Настасья выдана в оное же село за крестьянина
7	Жена члена семьи	У Илии жена Катерина Федорова старинная того же села
8	Дети (просто имя)	Михаил
		Мария
9	Сноха	У нее сноха вдова Ефимия Егорова дочь

5.1. Принципы распознавания типов

Если в ячейке первый элемент (слово) точно совпадает с каким-либо именем из словаря имен, тогда: если второй элемент является отчеством (производным от какого-либо мужского имени, оканчивающимся на «ов», «ев», «ин»: Иванов, Васильев, Вавилин), то запись в ячейке причисляем к **типу 1;** если второй элемент отсутствует либо не является отчеством, то запись причисляем к **типу 8.**

Если в ячейке первый элемент «Вдова» либо «Солдатка», то запись причисляем к **типу 1.**

Если в ячейке встречается элемент «сестра», то запись причисляем к **типу 4.**

Если в ячейке встречается элемент «брат», то запись причисляем к **типу 3.**

Если в ячейке встречается элемент «сноха», то запись причисляем к **типу 9.**

Если в ячейке встречается элемент «дочь» либо «дочери», то запись причисляем к **типу 6**.

Если в ячейке отсутствует элемент «дочь»/«дочери», но присутствует какой-либо из элементов: «дети», «сын», «сыновья», «рожденный(е) после ревизии», то запись причисляем к **типу 5**.

Если в ячейке присутствуют элементы «У него» и «жена», то запись причисляем к **типу 2**. Если вместо «него» стоит мужское имя в родительном падеже, персонаж с этим именем значится в одной из предшествующих записей этого же двора и относится к **типу 3, 5** либо **8**, то запись причисляем к **типу 7**. Если вместо «него» стоит мужское имя в родительном падеже, и в предшествующих

записях этого же двора записан единственный персонаж, относящийся к **типу 1**, то запись причисляем к **типу 2**.

5.2. Алгоритм заполнения реляционных таблиц

Нумерация персонажей была произведена на этапе формирования второй промежуточной таблицы. Дворы нумеруются последовательно по фактам выявления новых хозяев (записей **типа 1**).

Год рождения персонажа вычисляется путём вычитания из года проведения ревизии возраста персонажа на текущую перепись. В случае выбытия персонажа — путём вычитания из года предыдущей ревизии возраста на предыдущую перепись. Год и причина выбытия при этом фиксируются в таблице «Доп_характеристики».

Пол определяется на основании имени персонажа с помощью словаря имен (при этом производится контроль: для мужчин данные о возрасте должны присутствовать в колонках 4-й и 2-й промежуточной таблицы, для женщин — в 7-й и 5-й).

Прочие атрибуты персонажей (имя, отчество, фамилия, номер родственника, степень родства, характеристика) для каждого из типов определяются по следующим правилам:

Тип 1. Глава семьи.

Для мужчин: 1-й элемент – **имя** (из списка мужских имён), 2-й — **отчество**; **фамилия** (если есть) — ближайший из последующих элементов, кроме «сын», «вдов». Если имеется элемент «вдов», то он определяется как **характеристика**.

Для женщин-вдов, солдаток (глав семейств): 1-й элемент — **характеристика** («вдова», «солдатка»), 2-й — **имя**, 3-й — **отчество** (производное от мужского имени); **фамилия** (если есть) — ближайший из последующих элементов, кроме «дочь». Если есть фрагмент «старинная того же села», то он определяется как **характеристика**.

Номер родственника — Null, **степень родства** — Null.

Тип 2. Жена главы семьи.

Имя: следующий элемент после слова «жена», совпадающее с каким-либо именем из словаря женских имен.

Отчество: следующий элемент после имени (производное от мужского имени).

Фамилия: фамилия мужа (главы семьи).

Степень родства: жена.

Номер родственника: номер предыдущего персонажа.

Характеристика. Если в описании персонажа есть словосочетание «старинная того же села», то оно записывается в характеристики.

Тип 3. Брат главы семьи.

Имя. Берется тот элемент, который точно совпадает с каким-либо мужским именем в именительном падеже из словаря имен.

Отчество. Элемент после имени (производное от мужского имени).

Фамилия — фамилия главы семьи.

Степень родства: брат.

Номер родственника: номер первого персонажа, проживающего в данном дворе.

Характеристика. Если в описании персонажа есть элемент «вдов», то он записывается в характеристики.

Тип 4. Сестра главы семьи.

Имя. Берется тот элемент, который точно совпадает с каким-либо женским именем в именительном падеже из словаря имен.

Отчество. Элемент после имени (производное от мужского имени).

Фамилия (по мужу — при наличии): ближайший из последующих элементов, кроме «дочь».

Степень родства: Сестра.

Номер родственника: номер первого персонажа, проживающего в данном дворе.

Характеристика. Если в описании персонажа есть элементы «вдова», «солдатка», то они записываются в характеристики.

Тип 5. Дети (имя с предварительным текстом).

Если в описании есть фраза «написан(ы) в последней пред сей ревизии» либо «рожденный(ые) после ревизии», то **именем** является элемент, который стоит после слова «ревизии» и точно совпадает с каким-либо именем из словаря имен.

Если нет, то в качестве **имени** берем четвертый элемент от начала (после словосочетания «У них/него/нее дети/сын/сыновья»), сверив его со словарем имен.

Отчество. Элемент после имени, производное от мужского имени (указывалось только в случае, когда отец по какой-либо причине не записывался в число жителей данного двора). В случае его отсутствия идем вверх по таблице и, начиная с текущей строки, ищем первую запись, начинающуюся со слов «У них/него». Относительно этой записи берется ближайший из предшествующих персонажей (в пределах двора) с мужским именем (вместо элемента «него» может стоять мужское имя, тогда берется ближайший из предшествующих персонажей (в пределах двора) с этим именем). Номер этого персонажа будет **номером родственника**.

Фамилия: фамилия главы семьи.

Степень родства: сын или дочь (в зависимости от пола).

Номер родственника. Случай отсутствия отчества уже оговорен. В случае присутствия отчества идем вверх по таблице и, начиная с текущей строки, ищем первую запись, начинающуюся со слов «У нее». **Номер родственника** — номер предшествующего персонажа. Если вместо «нее» стоит женское имя, то берется номер ближайшего из предшествующих персонажей (в пределах двора) с этим именем.

Характеристика. Если в описании персонажа есть элемент «вдов», то он записывается в характеристики.

Тип 6. Дочь/дочери (специально выделенные в исходном документе).

Имя. Элемент, стоящий после слова «дочь/дочери» (должен совпадать с каким-либо женским именем из словаря имен). Если вместо имени стоит словосочетание «записана(ы) в последней пред сей ревизии», то именем является следующий элемент после этого словосочетания.

Отчество: так же, как для **типа 5**.

Фамилия: фамилия главы семьи.

Пол: женский.

Степень родства: дочь.

Номер родственника: так же, как для **типа 5**.

Характеристика. Если в описании персонажа упоминается факт выдачи замуж, то он фиксируется в таблице характеристик.

Тип 7. Жена члена семьи.

Имя. Берется элемент, стоящий после слова «жена» и точно совпадающий с каким-либо женским именем из словаря имен.

Отчество: элемент после имени (производное от мужского имени).

Фамилия: фамилия главы семьи.

Пол: женский.

Степень родства: жена.

Номер родственника. Номер ближайшего из предшествующих мужских персонажей (в пределах данного двора) с именем, которое в родительном падеже перед словом «жена».

Характеристика. Если есть, то «старинная того же села».

Тип 8. Дети (просто имя).

Имя. Берется единственный элемент, он должен точно совпадать с каким-либо именем из словаря имен.

Отчество: так же, как для **типа 5**. **Фамилия**: фамилия главы семьи.

Степень родства: сын или дочь (в зависимости от пола).

Номер родственника — так же, как для **типа 5**.

Характеристика. Если в описании персонажа есть элементы «вдов», «вдова» и пр. либо информация о выдаче замуж, то они записываются в характеристики.

Тип 9. Сноха (жена сына, отсутствующего в переписи).

После слов «сноха вдова», «сноха солдатка» и пр.: 1-й элемент — **имя**, 2-й – **отчество**. «Вдова», «солдатка» — **характеристики**. **Фамилия**: фамилия главы семьи. **Степень родства**: сноха.

Номер родственника. В случае присутствия элемента «У него/нее» — номер предшествующего персонажа, в случае присутствия элемента «У них» — номер предпредшествующего персонажа, в случае, если вместо «него/нее/них» стоит конкретное имя, то берется номер ближайшего из предшествующих персонажей (в пределах двора) с этим именем.

Если в описании персонажа имеется словосочетание «старинная того же села», то оно записывается в **характеристики**.

В результате применения описанного алгоритма, на основе изначальной табл. 1 будут сформированы следующие две реляционные табл. 3 и 4.

Таблица 3

Персонажи

Номер двора	**Номер персон.**	**Имя**	**Отчество**	**Фамилия**	**Пол**	**Год рожд.**	**Номер родств.**	**Степень родства**
1	1	Тихон	Герасимов	Null	муж.	1696	Null	Null
1	2	Анна	Максимова	Null	жен.	1703	1	жена
1	3	Овдотья	Герасимова	Null	жен.	1707	1	сестра
2	4	Михайло	Паникратов	Null	муж.	1705	Null	Null
2	5	Овдотья	Акинфиева	Null	жен.	1723	4	жена
2	6	Деомид	Михайлов	Null	муж	1733	4	сын
2	7	Прасковья	Михайлова	Null	жен.	1753	4	дочь
2	8	Овдотья	Михайлова	Null	жен.	1754	4	дочь
2	9	Прасковья	Григорьева	Null	жен.	1733	6	жена
2	10	Кондратей	Деомидов	Null	муж.	1754	6	сын
2	11	Ирина	Деомидова	Null	жен.	1764	6	дочь
2	12	Евфимия	Егорова	Null	жен.	1754	10	жена

Таблица 4

Доп_характеристики

Номер персон.	**Характеристика**
1	умре в 773 году
2	старинная того же села
2	умре в 774
3	вдова
3	умре в 770 году
5	старинная того же села
5	умре в 770
7	умре в 781
8	выдана в оное же село за крестьянина
9	старинная того же села
12	старинная того же села

Примечание. В исходном историческом документе могут встретиться записи, которые не относятся ни к одному из вышеперечисленных типов (например, во дворе проживает племянник хозяина двора — сын брата, оставшийся без родителей). Их количество, как правило, невелико, и соответствующие ячейки таблицы специальным образом маркируются для дальнейшего распознавания «вручную».

6. Заключение

Описан принцип и алгоритм автоматического выявления характеристик жителей какого-либо населенного пункта, зафиксированных в архивных документах переписей населения, на примере ревизских сказок 1782 и 1795 гг. В результате применения этого алгоритма формируются реляционные таблицы базы данных, позволяющие осуществлять автоматический поиск по заданным критериям. По такому же принципу подобные алгоритмы могут быть разработаны и для других ревизий, что позволит выстроить непрерывную цепочку исторических переписных ведомостей, переведенную в единый формат базы данных и охватывающую период с 1719 по 1858 годы. При наличии этой цепочки появляется возможность, начиная с конкретного предка, пройти «в глубь времен» и проследить историю своего рода, выявив при этом своих сородичей прошлых поколений и установив родовые связи. При этом возникает задача идентификации одного и того же персонажа в нескольких смежных ревизиях, которая может быть решена путем сопоставления имен, отчеств, возрастов и составов семей. Подробное решение этой задачи изложено в [7].

Литература

1. *Гарсиа-Молина Г., Ульман Д., Уидом Д.* Системы баз данных. Полный курс. — М.: Издательский дом «Вильямс», 2003. — 1088 с. [*Garsia-Molina G., Uljman D., Uidom D.* Sistemih baz dannihkh. Polnihyj kurs. — M.: Izdateljskiyj dom «Viljyams», 2003. — 1088 s.]
2. *Гринев М.* Системы управления полуструктурированными данными // Открытые системы. — 1999. — Т. 5–6. [*Grinev M.* Sistemih upravleniya polustrukturirovannihmi dannihmi // Otkrihtihe sistemih. — 1999. — T. 5–6.]

3. *Горелов С. С.* Эффективные модели поиска в базах полуструктурированных данных на основе иерархии схем документов. Диссертация на соискание степени кандидата физико-математических наук. — М., 2009. [*Gorelov S. S.* Ehffektivnihe modeli poiska v bazakh polustrukturirovannihkh dannihkh na osnove ierarkhii skhem dokumentov. Dissertaciya na soiskanie stepeni kandidata fiziko-matematicheskikh nauk. — M., 2009.]
4. ОБД Мемориал. — `http://www.obd-memorial.ru`. [OBD Memorial. — `http://www.obd-memorial.ru`.]
5. All Russia Family Tree. Российская генеалогия. — `http://www.vgd.ru`. [All Russia Family Tree. Rossiyjskaya genealogiya. — `http://www.vgd.ru`.]
6. АОТ. — `http://www.aot.ru`. [AOT. — `http://www.aot.ru`.]
7. *Брилева Н. Е.* Генерация родословных линий на базе оцифровки архивных документов переписи населения // Фестиваль науки в РУДН: Сборник работ студентов-победителей международных, всероссийских конкурсов, конференций, олимпиад. — РУДН, 2010. — С. 118–124. [*Brileva N. E.* Generaciya rodoslovnihkh liniyj na baze ocifrovki arkhivnihkh dokumentov perepisi naseleniya // Festivalj nauki v RUDN: Sbornik rabot studentov-pobediteleyj mezhdunarodnihkh, vserossiyjskikh konkursov, konferenciyj, olimpiad. — RUDN, 2010. — S. 118–124.]

UDC 004.652.4:929.52

Development of Relational Model of Data on the Inhabitants of the Russian Empire based on the Digitized Russian Census Documents

N. Ye. Brileva, A. S. Pankratov

Information Technologies Department
Peoples' Friendship University of Russia
Miklukho-Maklaya str., 6, Moscow, Russia, 117198

A technique of a database constructing which contains materials of the Russian Empire censuses, by the example of census returns of XVIII–XIX centuries is considered. The database is based on the digitization of archival materials census made in Excel-format. An algorithm for recognition of semantic structures in the records of census documents, the detection of attributes and rules for filling the relational tables are described. The algorithm is given for the specific census (4 and 5 revisions), but these algorithms can be also developed for other revisions and it will make possible to build a continuous chain of historical census lists, translated into a single database format. Such a chain can provide meaningful assistance in genealogical research.

Key words and phrases: database, semistructured data, digitizing, genealogy, census, census return.