Электронная коллекция биографических фактов

© Н.А.Маркова
Институт проблем Информатики РАН,
Москва
nMarkova@ipiran.ru

Аннотация

Предложен объектно-ориентированный метод формализации представления биографических данных, на основе которого построена модель хранения коллекции фактов. Метод учитывает динамику изменения атрибутов объектов и отношений между ними, рассчитан на работу с гетерогенными источниками, искаженными данными. Обсуждены перспективы реализации поддерживающего инструментария.

1 Введение

Биографические исследования – будь то изучение биографии выдающегося деятеля или группы лиц определенного круга; изыскания, проводимые в рамках изучения истории науки или краеведческие работы – обладают рядом общих задач. Необходимо собрать и обобщить биографические факты, под которым понимается «высказывание…, являющееся ответом на вопросы типа кто?, что?, когда?» [1] упорядочить их определенным образом, связать между собой и с объектами исторической реальности.

Материал, собираемый в рамках биографического исследования, а также результаты его анализа нужно как-то сохранять. Такое хранение, как правило, ведется бессистемно – в виде выписок на листках бумаги или в текстовых файлах. В то же время, существенная часть изучаемых данных формализуема: они имеют хронологическую привязку, отражают достаточно определенные социологические представления о характеристиках и связях изучаемых лиц. Все это может служить основой для систематизации хранения, и, в конечном счете, создания методики и инструментария, обеспечивающего автоматизацию процессов ввода, анализа, обобщения данных.

Три особенности биографической информации затрудняют построение их формальной модели. Во-первых, данные, характеризующие одних лиц, неприменимы к другим (или неизвестны для них). Во-вторых, названия тех или иных характеристик не только изменяются во времени, но и одновременно существуют в разных вариантах. Это относится к именам лиц, наименованиям организаций и географических объектов. Наконец, как это формулировал историк Л.Н.Гумилев в своих публичных лекциях: «источники все врут». Иначе, наличие дефектов в биографических данных, черпаемых из источников, скорее правило, чем исключение.

Представим метод формализации биографических данных, учитывающий перечисленные особенности. Определим основные проектные решения инструментария, автоматизирующего процессы ввода, хранения, организации доступа и анализа биографических данных. Реализуемый в настоящее время прототип такого инструментария рассчитан на работу индивидуального исследователя.

2 Формализация представления биографических данных

Основой формализации является всем знакомая анкета. Неанкетную – повествовательную часть биографии можно было бы существенно сократить, если бы относительно лиц, организаций, географических мест, связанных с основным лицом, тоже были бы составлены свои, пусть фрагментарные, анкеты, на которые можно было бы ссылаться.

Подавляющее большинство характеристик, фиксируемых в анкете, изменяются во времени. Местопребывание, род деятельности, семейное положение, а, возможно, и фамилия многократно меняются на протяжении жизни. То есть, для полноты картины там, где в обычной анкете мы ставим текущее или интегрированное значение, следует иметь таблицу, в которой изменяемые значения соотнесены со временем.

Существование (как минимум – значимость) характеристик и их возможные значения зависят от конкретно-исторической ситуации, следовательно, как набор «граф» анкеты, так и списки допустимых значений должны быть гибко настраиваемы.

Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.

Суммируя вышесказанное, определим формальное представление биографических данных как объектную модель, в которой объекты – люди и связанные с ними сущности – характеризуются атрибутами и связями. Причем, как объекты, так и атрибуты и связи хронологически определены (частный вид хронологической определенности – «датировка неизвестна»). Типы атрибутов и связей, а также их значения не закреплены, а определяются соответствующими «метками» на графе.

Биографическая информация далеко не исчерпывается формализуемой частью. Но и относительно данных, не подлежащих формализации, необходимо сохранять вполне определенную, идентифицирующую их информацию: источник, хронологические и географические рамки, сопряженные объекты.

Определим основные составляющие формальной модели биографических данных.

2.1 Объекты, атрибуты, отношения

Целевым объектом биографического исследования является человек или группа людей, принадлежащих к определенному кругу, информация о которых сохранилась в источниках. Изучаемое лицо связано с какими-то людьми, организациями, обществами, а также географическими и материальными объектами. В круг рассмотрения биографа включены документальные объекты – источники данных, а также документы, по отношению к которым изучаемое лицо является автором или адресатом. Обобщенная трактовка понятия «объект» позволяет применить ее и для таких категорий, как событие (например, «Великая отечественная война», в которой персонаж участвовал) или концепция (по отношению к которой персонаж автор или приверженец).

В конкретных исследованиях могут быть востребованы те или иные классы рассматриваемых объектов, под которыми будем понимать:

- Лица (индивидуумы, персоны, личности);
- Социальные объекты (семьи, организации, общества, группы);
- Географические объекты (места);
- Документальные объекты (документы, их совокупности и фрагменты);
- Материальные объекты (природные, технические, художественные);
- События (явления, процессы, активность, деятельность);
- Концепции (абстрактные понятия, идеи, области знания, дисциплины, методы, технологии).

Каждому объекту сопоставляется датировка – время жизни, существования. Датировка необходима также и для атрибутов объекта и для отношений между объектами. В такое-то время Иванов служил на Почтамте, тогда-то был женат на Петровой, тогда-то болел чахоткой.

Классы атрибутов и отношений зависят от классов определяемых (связываемых) объектов и также определяются спецификой исследований. Для связей лиц важнейшим классом является «Родство» со значениями («метками»): «Родитель», «Супруг» и т.п. Отношение между лицом и социальным объектом характеризуется классом «Позиция», под которым, в частном случае, понимается должность в учреждении. По отношению к учебному заведению лицо может выступать как «Преподаватель», «Ученик»; по отношению к клинике – «Пациент»; по отношению к научному обществу – «Член-корреспондент».

Необходимо допустить многозначность атрибутов и отношений, а также их взаимную зависимость. Например, один объект может обладать двумя разными атрибутами «Имя» с пересекающимися датировками: «Покрова Пресвятой Богородицы, что на Рву» и «Василия Блаженного». Не менее важно иметь возможность анализировать случайные и намеренные искажения биографических данных, содержащихся в разных источниках. Так отечественные документы учета в 18-м – начале 20-го веках фиксировали не год рождения, а возраст. При сопоставлении данных одного и того же лица за разные годы выявляются существенные расхождения. Часть из которых носит намеренный характер. Социальные статусы (в частности, возможность поступить в учебное заведение или занять должность) имели возрастные цензы, что служило причиной «поправления» возраста в нужную сторону.

2.2 Представление биографических фактов

Представим сведения об объекте формально, в виде совокупности фактов, фиксирующих модель изучаемого мира (конкретно-исторической ситуации) в терминах объект-атрибут-отношение. По ходу работы исследователь пополняет коллекцию фактов, как за счет интерпретации источников, так и в результате анализа – обобщения ранее накопленных данных. Определим основные виды фактов.

- Дефиниция объекта - утверждение о существовании объекта некоторого класса в определенный промежуток времени.
- Атрибут объекта - оценка значения атрибута объекта в определенный промежуток времени.
- Отношение объектов - констатация наличия определенного отношения между двумя объектами в определенный промежуток времени и, возможно, оценка его значения (метка на ребре графа).
- Связь фактов - высказывание, сопоставляющее факты логически, хронологически и т.п.

Дефиниция объекта является точкой привязки остальных фактов. Атрибут связан с одним объектом, отношение – с двумя (**Рис. 1**). Связь фактов соотносит любую пару в терминах «раньше/позже» или «следует/противоречит» и т.п.

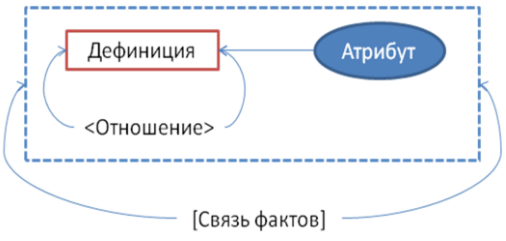

Рис. 1 Виды фактов

Определим информационное содержание фактов, – точнее, их записей в цифровой среде – в обобщенном виде.

Представление факта помечается идентификатором, для того, чтобы на него можно было бы сослаться. Для каждого факта фиксируется класс (соответственно, объекта, атрибута, отношения). Дополнительно, возможно, определяется некоторое уточняющее значение.

Дефиниция для удобства использования содержит (основное) имя объекта. Отдельные компоненты имени (для лица в отечественной традиции – личное имя, отчество, фамилия) представимы и в виде отдельных атрибутов, без чего не обойтись, в случае изменчивости имен на протяжении жизни лица.

Взаимная зависимость, содержательная избыточность сохраняемого набора фактов – важнейшая, практически значимая для биографических исследований черта. В случае с именами это особенно актуально. Один и тот же населенный пункт может многократно менять название или (что соответствует отечественной практике) иметь одновременно разные варианты наименований.

Представление факта должно содержать его датировку – период времени, в течение которого представленное фактом утверждение справедливо. В работе [2] была предложена универсальная форма представления датировки, в которой учтены все возможные сочетания хронологических сведений: точно определенный диапазон дат или оценки верхних/нижних границ начала/конца периода.

Факты определяются на основании интерпретации источников или выводятся путем умозаключения из других фактов. В любом случае, сохраняя факт, необходимо сохранить и ссылку на его происхождение - отражающий объект, в качестве которого могут выступать:

- Источник документ, вещественный объект, лицо-информатор;
- Интерпретатор – конечный исследователь («я») или предшествующее исследование, через документ–публикацию, что позволит проследить цепочку выводов от первоисточников;
- Автомат, выводящий факт из других фактов и правил вывода.

2.3 Логика биографических фактов

Помимо точных характеристик, как значений атрибутов и отношений, так и их датировки, в ходе исследования целесообразно фиксировать и то, что эти данные пока неизвестны. Прежде всего, речь идет об отсутствии данных и необходимости их выявления, то есть, о формировании исследовательского вопроса. Могут быть известны некоторые ограничения, тогда речь идет об уточнении. Наконец, при расхождениях в данных различных источников – что типично – ставится вопрос о выявлении корректного значения. Например, возраст Василия в «Исповедных ведомостях» 1743 года был – 1 год, в документе 1746 года ему 9 лет. Аппарат фактов может быть также использован для формулировки гипотез.

В работе [3] был предложен способ представления формализуемых биографических данных в виде логических формул, в которых помимо равенства для значений характеристик объектов применяются также неравенство, принадлежность подмножествам, а для упорядоченных значений еще и больше/меньше.

При определении связей между фактами применяются категории «раньше/позже», «причина/следствие». Наконец факт, как логическое утверждение, подлежит оценке, которая может иметь как точную (ИСТИНА/ЛОЖЬ), так и промежуточную оценку правдоподобия в виде числа в диапазоне от 0 (ЛОЖЬ) до 1 (ИСТИНА).

Представление биографических фактов в предложенном виде позволит интегрировать данные, получаемые из разных источников, проверять непротиворечивость, интерполировать, корректно ставить новые исследовательские вопросы.

2.4. Другие модели представления биографических данных

Готовых концептуальных моделей, в полном объеме отражающих специфику биографических исследований, учитывающих темпоральные зависимости, наличие искажений, разнородность источников, вариативность названий – все «неудобные» для реализации особенности исторических изысканий, найти не удалось.

В наиболее близкой к рассматриваемой проблематике отрасли исторической информатики – просопографических базах данных – задачи построения обобщенной объектной модели, судя по публикациям, не ставятся. Реализации рассчитаны на конкретный вид исследований и/или на конкретный круг источников.

Стандартом де-факто для представления биографической информации при обмене данными между генеалогическими программами является давно устаревшая модель Genealogical Data Communications (GEDCOM). Усовершенствованный вариант GEDCOM 6.0, основанный на xml, [4] был выпущен в 2002 году, но до сих пор, фактически, никем не используется. Конкурентом GEDCOM, тоже определяемым, как спецификация формата обмена, является стандарт GenXML[5]. Помимо

более строгой, структурной упорядоченности, он несет в себе несколько принципиально новых положений, отражающих практику биографических исследований. В частности, в нем явным образом определяется процесс исследования: введены понятия «свидетельство» и «заключение». Но главное, GenXML открыт для добавления новых типов атрибутов и событий. К сожалению, ни в GEDCOM, ни в GenXML не отражены важнейшие свойства биографической информации: временная изменчивость и взаимная зависимость характеристик.

Важным шагом в сторону эффективного представления биографических сведений является стандарт «Функциональных требований к авторитетным данным» – FRAD [6]. Сопоставим FRAD с предлагаемой в данной работе моделью. (Учтем терминологические расхождения: во FRAD под «объектами» понимаются абстрактные категории, которые в настоящей работе названы «классами объектов»; соответственно, «пример объекта» FRAD соответствует «объекту» в нашем рассмотрении).

В целом, модели весьма схожи. На уровне концепции сущность-атрибут-отношение значимым различием является то, что для FRAD «имя» - это объект, в нашем рассмотрении имя – это атрибут. Авторы FRAD справедливо замечают, что нечто интерпретируется как атрибут или объект в зависимости от использования. А для систем каталогизации, на которые, в основном, рассчитаны FRAD, традиционно центральной сущностью является «имя». В биографической модели целесообразно рассматривать имя, как один из атрибутов, многозначный и изменяемый во времени. Что покрывает множество вводимых FRAD понятий («псевдонимы», «духовные», «светские» имена, связь прежнее/ более позднее имя и др.). Кроме того, учет вариативности компонентов имени (личного имени, отчества) целесообразно предусматривать для всех возможных их использований, а не только для конкретного имени конкретного лица.

Принципиальным преимуществом предлагаемой биографической модели перед FRAD для задач исторических реконструкций является ее «историзм»: всем объектам, атрибутам, связям сопоставлены хронологические рамки. Другая важная черта биографической модели, чрезвычайно существенная для представления данных в процессе исследования (а не конечных результатов, как FRAD) – это возможность отражения оценки правдоподобия сведений, их логической связности, наличия противоречий - логики биографических фактов.

3. Основные проектные решения

Рассмотрим архитектуру инструментария, поддерживающего процессы сбора, хранения и анализа биографических данных (Рисунок 2). В качестве модели представления данных будем

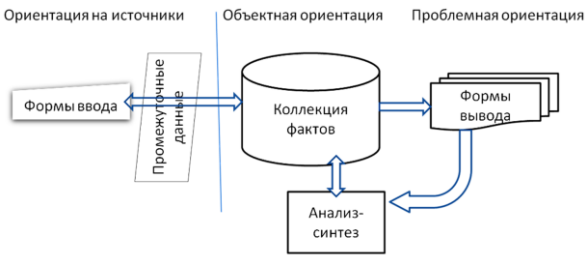

Рис. 2 Архитектура биографического инструментария

использовать предложенные решения по формулировке биографических фактов. Ограничимся сферой деятельности индивидуального исследователя, то есть на текущем этапе не будем рассматривать проблемы взаимодействия в группе пользователей, а также вопросы работы с крупномасштабными хранилищами данных.

Центральный компонент инструментария – хранилище – содержит коллекцию фактов, к которым помимо собственно исторических сведений отнесем данные о происхождении этих сведений – библиографическую и археографическую информацию. Информация о существовании некоторого документа, времени его создания, авторства, места хранения – по сути, и является набором исторических фактов, и, как мы уже констатировали, документ – это один из классов исследуемых объектов.

Коллекция фактов, в которую оперативно интегрируются данные из различных источников, представляет единый ресурс биографического исследования - «виртуальный метаисточник» по терминологии специалистов в области исторической информатики [7].

Специфика биографического исследования, особенно, если изучаются лица из прошлого, документов о которых сохранилось немного, состоит в том, что невозможно заранее отсечь «ненужную» информацию. Изучая некоторый источник, например, список выпускников учебного заведения, и найдя интересующее нас лицо, целесообразно сохранять не только факт его обучения там-то в такие-то годы, но и сведения о его однокашниках. Впоследствии из «их» документов, вполне вероятно, можно будет почерпнуть данные, если не непосредственно об изучаемом лице, то об окружающей его среде. Коллекция фактов – это ресурс, независящий ни от конкретных источников, ни от текущих исследовательских задач. Его содержание – все возможные характеристики объектов исследования и их окружения, сохраняемые, возможно, «про запас», в расчете на будущие исследовательские вопросы.

В то же время, ввод данных, по возможности, должен соответствовать конкретному документу-источнику, а анализ целям конкретного исследования. В первом случае необходимо опираться на источнико-ориентированную, во втором на проблемно-ориентированную информационную модель. Вопрос о приоритете того или иного подхода

является одним из основных направлений дискуссий в исторической информатике [8]. Опора на предложенную объектно-ориентированную формальную модель позволяет разрешить противоречия между подходами. Трансформация представлений между моделями, при условии, что центральной является объектно-ориентированная, – несложная техническая задача, решение которой позволит реализовать максимальную наглядность данных на разных этапах работы исследователя. При вводе пользователь будет видеть визуальный аналог изучаемого источника, при анализе фактов – структурные формы (таблицы, схемы), в которых вычленяются изучаемые в конкретном исследовании аспекты. В качестве иллюстрации рассмотрим представление в терминах формальной модели биографических фактов данных, извлекаемых из типового источника.

4. Пример интерпретации данных источника

То, что предложенный формализм, столь обобщенно обращающийся с объектами разного рода, пригоден для представления содержания исторического источника, покажем на примере интерпретации одних из наиболее значимых источников формальных биографических сведений России 18 - 19 веков – «Исповедных ведомостей».

Ведомость представляла собой таблицу, заполняемую причтом конкретной церкви, перечисляющую всех жителей прихода. Данные о каждом лице соответствовали графам таблицы, и включали: номер дома, пол, «звание», возраст и собственно «показание действа», то есть, исповедовался ли, и если нет, то по какой причине.

«Звание» включало имя (фамилию, личное имя отчество), сведения о роде занятий и статусе, а также о родственных отношениях с проживающими вместе лицами.

Представим данные из «Исповедной ведомости» в терминах предлагаемой формальной модели (Рисунок 3).

Рассматриваемыми объектами являются: лица, места (географические местности), приходская церковь и сам документ – «Исповедная ведомость». Кроме того могут быть указаны места службы упоминаемых лиц или/и их социальные статусы. Для лиц указываются атрибуты – пол и возраст (на момент фиксации). Между лицами определены родственные отношения. Кроме того фиксированы отношения местопребывания с точностью до дома, входящего в некоторое географическое образование.

Тонкие стрелки показывают связь фактов с объектом отражения.

Представленные таким образом сведения носят объектно-ориентированный характер и уже не зависят от источника, что позволяет сопоставлять их с фактами, полученными из других источников, возможно, другого вида. При этом мы можем

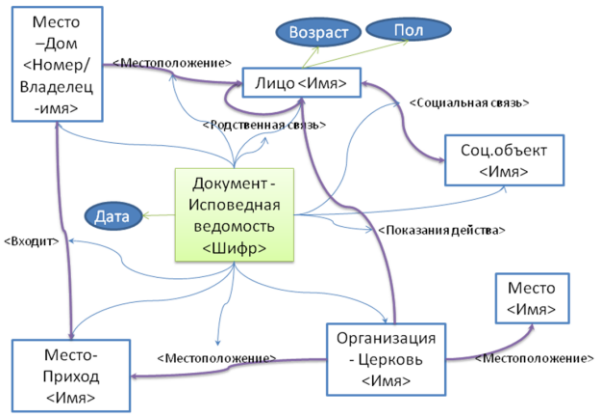

Рис. 3 Концептуальная модель *данных из «Исповедных ведомостей»*

привлечь автомат, например, для вычисления года рождения по возрасту и дате фиксации.

5. Нормали биографического исследования

Биографические факты взаимозависимы. В ограниченной хронологическими, географическими, предметными рамками сфере конкретного исследования действует конкретный набор правил регламентирующих эти зависимости. Такие правила будем называть *Нормалями*. К нормалям отнесем словари классов и возможных значений фактов, ограничения на их сочетания, синонимию имен и их компонентов и т.п. Сведения о структуре и правилах интерпретации источников, а также шаблоны выводных форм, используемых для анализа накопленных фактов, являются нормалями.

Нормали представляют формулировки законов природы (например, ограничения на разницу между возрастом родителей и ребенка), юридические нормы, уставы организаций, обобщения ранее накопленных фактов. Типичной нормалью является список должностей некоторого предприятия. Служебные отношения лиц с данным предприятием определяются выбором из этого списка. Другой пример нормали – список соответствий имя отца–отчество. Нормаль – список личных имен-синонимов поможет идентифицировать лица при вариативности их имен. Например, личное имя «Иван» могло быть зафиксировано, как «Иоанн»; «Акулина» – «Акилина»; «Ксения» – «Аксинья»; «Георгий» – «Егор», «Егорий», «Юрий».

Так же, как и конкретные факты, нормали и их компоненты хронологически определены. Наличие той или иной должности или звания ограничено конкретным временным промежутком. Например, «Табель о рангах» применялась с 1722 по 1917 гг., а входящий в нее чин «Сенатский регистратор» с 1764 по 1834.

Сложностью правила, формулируемого нормалью, определяется возможная форма ее подключения к коллекции фактов. В простейшем случае нормаль – это словарь значений. Его использование способствует автоматизации ввода данных. Вместо

набора текста исследователь выбирает значение из списка допустимых в данном контексте значений. Более сложные правила, фиксируемые нормалями, либо подлежат анализу/обработке с помощью дополнительно вводимых программных модулей, либо представляют «памятки» для ручного контроля исследователем.

Структура и правила интерпретации источника определенного вида – это и структурная организация «родового» документального объекта (своего рода шаблон), и правила интерпретации данных конкретного источника. С точки зрения вводящего данные исследователя было бы удобно, чтобы отображение архивного источника выглядело максимально «близко к тексту». Рационально заполнять форму с графами, соответствующими графам интерпретируемого документа. При этом там, где это возможно, автомат может учесть типовые последовательности ввода и словарные ограничения, существенно сократив время занесения данных и повысив его надежность.

В общей структуре инструментария биографического исследования нормали занимают промежуточное положение между универсальной формой представления коллекции фактов и ее специализированным, предназначенным для решения конкретного комплекса исследовательских задач наполнением.

Повышение эффективности работы исследователя, достигаемое при использовании автоматизированных средств отслеживания правил, фиксируемых нормалями, приходится сопоставлять с трудоемкостью разработки этих средств.

6. Заключение

Предложенную модель представления биографических данных отличают следующие черты:

- рассматриваются не только лица, но и объекты иной природы вместе с их структурой и историей;
- помимо атрибутов объекта анализируются характеристики отношений между объектами;
- все характеристики рассматриваются в динамике изменения их значений;
- наличие характеристик, их возможные значения и взаимозависимости определяются конкретно историческими знаниями – нормалями, которые также изменяются во времени;
- для представления совокупности установленных фактов, намеченных к рассмотрению вопросов, а также гипотез предложена единая форма.

Представление биографических сведений в виде коллекции фактов открывает новые возможности для эффективного хранения, поиска, анализа и интеграции данных. В настоящее время осуществляется прототипная реализация соответствующего инструментария. В качестве среды хранения выбрана СУБД MS Access, что не является принципиальным – подошла бы любая реляционная база, а обусловлено ее доступностью. План реализации предусматривает формирование унифицированных средств, поддерживающих функционирование ядра коллекции фактов, а также формирование нормалей и пробные наполнения коллекции для двух видов биографических исследований.

Первый вид – исследование биографии ученого, на основе изучения архивных документов и публикаций, касающихся как его самого, так и лиц и организаций, с ним связанных. Второй вид исследования – просопографическое – объединение данных, полученных из разнообразных документов церковно-приходского учета, в территориально-ограниченной области для нескольких поколений родственных семей. Во втором случае предстоит разработать формы ввода для типовых видов учетных документов. Оценка масштаба коллекции фактов для выбранных примеров исследований дает сотни лиц (если изучается биография ученого – его коллеги, учителя, ученики, родня), и десятках тысяч фактов.

Пробная эксплуатация инструментария должна выявить проблемы и перспективы его дальнейшего развития. Вполне вероятно, что предложенный подход может быть востребован не только для изучения биографий, но и для других конкретных исторических дисциплин: истории отрасли, организации, общества, края, где фокусом внимания будут не лица, а объекты другой природы.

Литература

[1] В. Л. Валевский. Биографика как дисциплина гуманитарного цикла // Лица: биографический альманах. - СПб. : Феникс, 1995. - Вып. 6. - С.33-68.

[2] Н. А. Маркова. Формализация представления биографических данных: рабочее поле биографического исследования //Системы и средства информатики, 2011, вып.21:2, С. 162–170.

[3] Н. А. Маркова. Логика биографических фактов //Информатика и её применения, 2012, Т. 6. Вып. 2. С. 49–58.

[4] GEDCOM XML Specification, Release 6.0. http://xml.coverpages.org/Gedcom-XMLv60.pdf

[5] GenXML 3.0 16.06.2010. http://www.cosoft.org/genxml/GenXML30.pdf

[6] Функциональные требования к авторитетным данным. Концептуальная модель: рабочая группа ИФЛА по разработке функциональных требований к авторитетным записям и их нумерации (FRANAR): заключительный отчет, декабрь 2008 / Междунар. Федерация библ. ассоц. и учреждений, Рос. библ. ассоц. ; под ред. Г. Е. Паттона. - СПб. : Российская национальная библиотека, 2011. - 115 с.

[7] Ю.Ю. Юмашева. Историография просопографии //Известия уральского государственного университета. - Екатеринбург : № 39.- 2005. - Гуманитарные науки. Вып. 10. С. 95-127.

[8] Проблемно-ориентированный и источнико-ориентированный подход – противоречие или синтез. http://www.yartel.ru/old/stat/pdisscas.html

Digital Collection of Biographic Facts

Natalia A. Markova

This article suggests an object-oriented method for formalizing biographical data. The method serves as a base for the model of a collection of facts repository. The method takes into account the dynamics of change attributes of objects and relationships between them. It is designed to operate with data (that may be faulty) obtained from heterogeneous sources. The article discusses prospects for the implementation of supporting tools.