Ю. Ю. Юмашева

МЕТАИСТОЧНИК: К ВОПРОСУ О ВЕРИФИЦИРУЕМОСТИ ДАННЫХ

За несколько десятилетий своего существования историческая информатика — сравнительно молодое направление в исторической науке – накопила значительный арсенал компьютерных методов и методик, позволяющих эффективно проводить работу с массовыми источниками и решать сложные конкретно-исторические проблемы. Основным достоинством этих методов является их «прозрачность», позволяющая легко верифицировать полученные в ходе проведенного исследования результаты и сделанные на их основании выводы, а в случае необходимости — полностью повторить научный эксперимент.

Однако эти бесспорные преимущества могут быть сведены к нулю из-за особенностей «докомпьютерного» этапа работы с массовыми историческими источниками. Свой особый отпечаток накладывает специфика различных видов и типов исторических материалов, принципы их отбора, оценки достоверности, формирования выборок, создания виртуального метаисточника и т. п. Каждый из перечисленных факторов может стать тем подводным камнем, о который рискует разбиться построенный признанными мастерами по превосходным чертежам корабль научного исследования.

Отечественная историография применения количественных методов обладает обширным опытом решения обозначенных проблем как на теоретико-методологическом уровне, так и в рамках конкретно-исторических исследований. В данной статье мы не будем останавливаться на подробном рассмотрении всего спектра предлагаемых решений, а сконцентрируем внимание на одной из наиболее часто встречающихся задач при работе с разными видами массовых материалов — задаче формирования тематического комплекса источников, проверки их достоверности, верификации сведений, учета разночтений и принципов отбора информации при создании нового единого «метаисточника», в который «сводится» отобранная информация и который в дальнейшем подвергается математическому анализу.

Нетрудно догадаться, что в описанном методе работы ключевым моментом становятся критерии отбора и оценки достоверности информа-

ции. Включение в метаисточник недостоверных сведений приводит к неправильным статистическим результатам и сделанным на их основе выводам. При этом в существующей исследовательской практике отсутствует механизм, позволяющий проверять правильность результатов и устойчивость выводов, изменяя исходные сведения, на которых они основаны.

В силу вышесказанного этап формирования метаисточника является, пожалуй, самым притягательным для критики моментом исследования. Действительно, в исторической информатике — направлении, где каждый шаг исследователя в решении той или иной конкретно-исторической задачи имеет не только логическое основание, но и математическое подтверждение, именно этот этап является наиболее «непрозрачным» для стороннего наблюдателя, именно его невозможно проверить опытным путем.

Этот факт особенно печален, поскольку создание метаисточника является вторым по популярности методом при работе с разновидовыми (нарративными и неструктурированными) документами при создании реляционных баз данных. В настоящее время без него не обходится практически ни одно исследование. Однако в публикациях-отчетах об этих работах описание технологии создания метаисточника либо отсутствует, либо заменяется невнятным упоминанием о том, что такой этап был осуществлен.

Думается, что эта «фигура молчания» отнюдь не случайна. К сожалению, в современной историографии по данному вопросу не предложено ни одного более или менее приемлемого алгоритма, позволяющего осуществлять решение этой задачи верифицируемыми компьютерными методами. И на практике вся работа по выбору сведений из источников, оценке их достоверности и приему окончательного решения о включении в метаисточник проводится «в уме» конкретного исследователя без возможности ее точного повторения и воспроизведения.

В предлагаемой статье на примере исследовательского проекта «Иноземцы на русской службе в первой трети XVII в.», в котором автор выступала в качестве «идеолога» компьютерной реализации, хотелось бы представить один из возможных «компьютерных» вариантов решения обозначенной проблемы.

Проект «Иноземцы на русской службе в первой трети XVII в.» был начат воронежскими исследователями О. В. Скобелкиным и В. И. Бесединым в конце 1990-х гг. Постановка исследовательской задачи предполагала создание просопографической базы данных[1] и дальнейшее изучение коллективной биографии иноземцев, прибывших на русскую службу в начала XVII в. В качестве источниковой базы были привлечены документы, отложившиеся во многочисленных фондах Российского государственного архива древних актов (РГАДА)[2].

Таким образом, уже в самом начале исследования в ходе архивных изысканий создавалась своеобразная «виртуальная» коллекция источников, объединенная одной общей проблемой — упоминанием служилых иноземцев.

Всего было выявлено около 300 единиц хранения, имеющих отношение к изучаемому объекту. Это разновидовые источники, не обладающие ни единством или формулярностью формы, ни однотипностью содержания, ни структурированностью содержащейся в них информации. Чаще всего это были нарративные документы или делопроизводственная документация регистрационно-учетного, информационно-отчетного или просительного характера.

Основным критерием отбора этих материалов было то, что они содержали ту или иную информацию об иноземцах, прибывших на русскую службу в первой четверти XVII в. При этом даже «количество» искомой информации в каждом из источников колебалось от простого упоминания до развернутого повествования на десятках и даже сотнях листов. Особо необходимо отметить то, что только незначительная часть этих материалов введена в научный оборот в единичных исследованиях и немногочисленных публикациях.

Все вышесказанное потребовало выработки нестандартных подходов и алгоритмов к решению типологических задач подготовки и переноса информации источников в структуру базы данных, т. е. в единый вновь создаваемый виртуальный метаисточник.

В ходе дальнейшей работы выяснилось, что выявленный комплекс содержит в себе все возможные варианты соотношения «источник – изучаемый объект»:

- один источник — одна персоналия;
- несколько источников — несколько персоналий;
- несколько источников — одна персоналия;
- один источник — несколько персоналий.

Превалирующими в выявленном комплексе были источники, представляющие собой первый и третий из описанных вариантов. Работа с первым вариантом с точки зрения создания баз данных является наиболее оптимальным и хорошо отработанным способом. Третий же вариант представляет наибольшую сложность для исследователей, так как для каждой персоналии может быть собственное количество источников (более двух), сведения которых могут как дополнять и детализировать, так и опровергать друг друга. Кроме того, этот вариант работы с источниками требует от исследователя виртуозного владения традиционными методами источниковедческого анализа в части отбора, оценки и верификации информации, содержащейся в различных источниках, т. к. результатом

осуществления этой подготовительной работы будет создание мета-источника. Конечно, было бы идеально, если бы при этом удалось сохранить и все «исходные» данные для последующего вероятного использования (реализация источниково-ориентированного подхода к созданию базы данных).

К сожалению, как уже было отмечено в самом начале статьи, приходится признать, что среди многочисленных исследований, выполненных к настоящему времени на основе варианта соотношения «несколько источников – один объект исследования», нет ни одной работы, где полностью раскрывался бы инструментарий проведения подобной источниковедческой работы с сохранением исходных материалов или их повторным использованием на основе уже реализованных исследований. Иными словами, первичные данные источниковедческого анализа оказывались невостребованными ни самими авторами исследований, ни их последователями, разрабатывающими аналогичную тематику. Такое пренебрежение к уже обработанным материалам объясняется наличием огромных массивов пока еще не введенных в научный оборот и не изученных источников и, кроме того, отсутствием разработанных и опробованных методик повторного изучения уже исследованных кем-то документов. То, что данные материалы могут содержать латентную информацию, в корне меняющую уже сделанные выводы, как-то упускается из виду.

Эта нерешенная методическая и технологическая проблема в нашем исследовании приобретает особое звучание в связи с тем, что привлеченные к работе источники в известном смысле уникальны, доступ к ним исследователей затруднен, а прочтение скорописных текстов XVII в. и интерпретация их информации чрезвычайно сложны.

Учитывая это обстоятельство, было принято решение разработать такой алгоритм работы с источниками, который бы позволял:

– полностью сохранить их информацию в базе данных с целью ввода в постоянный научный оборот выявленного массива документов;

– создать такую структуру базы данных, которая бы позволила зафиксировать и точно локализовать все разночтения, которые имеются в источниках, проводить «прозрачную» верификацию достоверности и гибкий отбор представленной информации;

– применить весь комплекс исследовательских методик для формирования коллективной биографии изучаемых персонажей.

Попытки решения первой из поставленных задач предлагались неоднократно. Так, в середине 1990-х гг. исследователь из Нидерландов Л. Брере разработал дополнительный модуль к СУБД dBASE IV (программа Socrates), позволявший представить в структуре одной записи базы данных небольшой фрагмент источника в виде машиночитаемого

текста и заполненные извлеченной из него информацией поля[3]. К сожалению, этот подход не получил широкого распространения по двум очевидным причинам. Первая заключалась в том, что работа по предложенному алгоритму в силу необходимости подготовки машиночитаемых вариантов источников значительно замедлялась, а вторая — в том, что разработанный метод обеспечивал обработку информации только в случае «один источник — один объект исследования».

Тем не менее в рамках представляемой работы опыт Л. Брере стал отправной точкой в поиске более совершенного механизма решения насущной задачи, который к тому же должен был обеспечить и решение второй проблемы — фиксацию в базе данных всех разночтений в привеченных материалах с точным указанием на источник их происхождения.

Выработанный в ходе исследования алгоритм оказался довольно простым, но эффективным. На первом этапе после внимательного ознакомления с источниками было принято решение создать таблицу «Источник», в которой бы содержалось краткое описание документа, осуществленное по усеченному стандарту описания, принятому в архивной практике (архив, фонд, опись, лист, лицевая сторона/оборот, автор), полный машиночитаемый текст и образ документа (если таковой можно было достать). Эта работа, даже если бы ее результаты впоследствии не могли быть интегрированы в создаваемую систему, не была бессмысленной, т. к. позволяла провести своеобразную «инвентаризацию» виртуальной коллекции источников: точно определить их количество, объемы текста, распределение по фондам, типам и видам, выявить цифру уже опубликованных и впервые вводящихся в научный оборот, провести первичное осмысление и простые статистические подсчеты. Более того, созданная таблица обладала самостоятельной ценностью[4], т. к. полностью подготавливала выявленные и представленные в электронном виде источники к использованию, минуя обращение последующих исследователей в архив. Фактически созданная таблица представляет собой полнотекстовую базу данных, в которой осуществлена электронная публикация источников в соответствии с принятыми в археографии требованиями.

На втором этапе шла разработка структуры реляционной базы данных, в которой бы отражалась информация, почерпнутая в выявленных источниках о каждом из исследуемых персонажей. Этот этап, состоящий из двух подэтапов (инфологического и даталогического проектирования), не представляет особой трудности и является стандартным при разработке баз данных[5]. Его результатом стало создание системы, состоящей из одной основной (главной), 16 вспомогательных и 26 дополнительных (таблицы подстановок) таблиц.

В соответствии с поставленной задачей во вспомогательные таблицы были выделены следующие аспекты (атрибуты):

1) связанные со службой: выход (переход на русскую службу), основные этапы службы (с особой регистрацией случаев участия в боевых действиях), категории иноземчества, оклад;

2) имущественные вопросы: жалованье, сведения о зависимых людях, собственном хозяйстве, челобитные о задержке жалованья и резолюции на них;

3) информация о семье — женах, детях и других упоминаемых родственниках.

Особую сложность при проектировании представляли такие темы, как жалованье и основные этапы службы. Формализация разнообразных по форме, составу, периодичности, причинности, размерам и материальному выражению выплат составила одну из самых трудноразрешимых задач. Не меньшей проблемой была и логика формализации сведений о боевом пути иноземца. Довольно запутанная система прохождения военной службы начала XVII в., нестабильная ситуация в стране, фрагментарно отразившиеся в привлеченных исторических документах, накладывали существенный отпечаток на разработку структуры соответствующей таблицы. Именно поэтому в каждой из упомянутых вспомогательных таблиц предусмотрено поле «Примечание», в которое выносятся уникальные факты, не поддающиеся формализации и не представляющие интереса в дальнейшем с точки зрения статистических процедур.

В самостоятельные таблицы были выделена информация об имени и фамилии персонажа, написанных на родном для него языке, а также о разночтениях в их написании при переводе на русский, выявленных в российских документах.

Отдельной вспомогательной таблицей стала таблица «Источники-1», в которой фиксировались все документы, где имелась какая-либо информация о конкретном персонаже. Для того чтобы не переписывать в ней каждый раз заново все сведения об использованных источниках, был применен стандартный механизм поля с типом данных «Подстановка», где в качестве подстановки выступал определенный набор полей, уникально характеризующий каждый из документов таблицы «Источник». При этом, с одной стороны, легко выявить, в каких конкретно источниках упоминался тот или иной персонаж, а с другой — понять, как часто тот или иной источник привлекался к работе.

Этот простой прием натолкнул на мысль о более широком применении данного подхода. В каждой вспомогательной таблице были введены собственное поле «Источник» с подстановкой из упомянутой таблицы и логическое поле «Достоверность». Это, в свою очередь, кардинально из-

менило само существо вспомогательных таблиц. Традиционно они использовались для отражения неповторяющихся, динамических или иных изменяющихся, например во времени, сведений, уже прошедших на этапе «докомпьютерной» работы с источниками исследовательский отбор и верификацию. К примеру, если таблица посвящена такой характеристике (атрибуту), как основные этапы службы, то каждая ее запись отражает определенную ступеньку карьерного роста. При этом если сведения для данной таблицы собирались из нескольких источников, то полностью дублирующаяся в них информация в базу вносится один раз, а из нескольких противоречивых вариантов на основе некоторых критериев отбора исследователем выбирается один, который и вносится в базу. Как правило, и в первом, и во втором случае ссылка на источник отсутствует.

Таким образом, формируется набор записей, лишенный повторов информации и полностью готовый для статистической обработки. В конечном итоге из таких наборов записей и формируется метаисточник. Если возникает необходимость верифицировать правильность и правомерность включения конкретной информации в этот набор (или метаисточник в целом), то нужно вновь поднять все источники и провести источниковедческую работу заново, выявляя в них интересующие сведения, отбрасывая дубли и отбирая одно значение из энного количества спорных.

В предложенном решении все несколько иначе. Во вспомогательную таблицу вносятся *все* данные, касающиеся изучаемого атрибута, и столько раз, сколько они возникают в источниках. Таким образом, в наборе записей существует и уникальная информация, и дублирующаяся, и все разночтения и противоречия. При этом каждая запись всегда имеет ссылку на источник. Отбор нужных для формирования метаисточника записей осуществляется с помощью установки маркера в поле «Достоверность». В этом случае исследователь отмечает уникальные записи, любую одну из дублирующихся и ту из противоречивых, которую считает наиболее достоверной. Результатом осуществления простого запроса с условием «Истина» в поле «Достоверность» будет искомый набор записей, готовый к включению в метаисточник.

Таким образом в базе данных сохраняется *вся информация об изучаемой предметной области с жесткой локализацией и связью с конкретным источником*, а метаисточник формируется на основе этой базы методом запросов.

Описанный механизм работы позволяет проводить «прозрачную» верификацию включения сведений в метаисточник, а при проведении статистических процедур и интерпретации полученных данных осуществлять их проверку на «устойчивость»[6] выводов.

Изменение существа вспомогательных таблиц и базы в целом повлекло за собой и изменение последовательности работы с ней: сначала

идет заполнение таблицы «Источник» и только затем — полей основной и вспомогательных таблиц. При этом на экране компьютера постоянно присутствуют две формы: форма из таблицы «Источник» с машиночитаемым текстом или образом конкретного документа и форма основной таблицы (с интегрированными в нее вспомогательными), в которую переносятся (иногда методом копирования) сведения из представленного документа.

Постоянное присутствие текста источника (источников) на экране значительно повышает верифицируемость отобранных для полей базы данных сведений, дает возможность оценить их в общем контексте документа, а также принять решение об оперативном изменении набора полей базы данных (добавлении или изъятии конкретных тематических полей), т. е. придает всей структуре базы определенную гибкость.

Безусловно, данный подход не лишен недостатков, главным из которых является фактор времени. Действительно, скорость создания такой базы значительно замедляется, во-первых, за счет временных затрат на подготовку электронной версии привлеченных документов (в виде образов или машиночитаемого текста), во-вторых, за счет необходимости отражения в базе всех (и уникальных, и дублирующихся, и противоречивых) сведений[7]. Кроме того, база данных значительно «утяжеляется» в силу увеличения количества записей во вспомогательных таблицах и необходимости хранения объемных фрагментов текста или изображений. Однако два последних возражения могут быть сняты благодаря наличию довольно совершенных программных продуктов. К примеру, в Государственном историческом музее при создании служебных баз данных учетной документации опытным путем доказано, что даже стандартная СУБД Access, представленная в пакете MS Office, имеет возможность работать с реляционными базами, насчитывающими около 1 млн записей. При переходе к более мощным СУБД или языкам программирования (Oracle, Lunix, SQL и т. п.) эта проблема снимается вовсе.

Вопрос о хранении больших объемов текста или образов также имеет давно проверенное решение: в базе данных представляется не сам текст или образ, а гиперссылка, открывающая файл необходимого документа, который может храниться на другом носителе отдельно от базы. Этот прием широко известен и распространен. В частности, на нем построены все имиджинговые системы, функционирующие в оборудованных компьютерных читальных залах крупнейших библиотек, архивов и музеев мира.

Таким образом, единственным существенным недостатком остаются временные затраты на создание подобной базы. Однако думается, что ввод в постоянный научный оборот источников с возможностью их многократного повторного использования без ущерба для оригинала оправ-

дывает эти трудозатраты, равно как и повышение верифицируемости результатов научных изысканий.

Описанный алгоритм не является чем-то уникальным и неповторимым. Его апробация была проведена как в рамках представляемой работы по иноземцам на русской службе XVII в., так и на практических занятиях по курсу «Базы и банки данных в исторических исследованиях» со студентами II курса факультета технотронных архивов и документов Историко-архивного института РГГУ. Сравнительная простота логики построения структуры базы дает возможность активно использовать данный инструментарий в широкой исследовательской практике.

---

[1] Подробнее о просопографии см.: *Юмашева Ю. Ю.* Историография просопографии // Изв. Урал. гос. ун-та. [Сер.] Гуманитарные науки. Вып. 10. 2005. № 39. С. 95–127.

[2] См., например: *О. В. Скобелкин, В. И. Беседин* Иностранцы на русской службе в царствование Михаила Федоровича // Информ. бюлл. Ассоциации «История и компьютер». 1997. № 21, март. С. 34–36; *Скобелкин О. В.* Списки служилых иноземцев 1-й половины XVII в. как источник по истории русского войска // Армия в истории России. Курск, 1997. С. 14–17; *Он же.* Организация службы иноземцев на защите южных рубежей России в первой половине XVII в. // Население и территория Центрального Черноземья и Запада России в прошлом и настоящем. Воронеж, 2000. С. 19–21; *Он же.* Тульская служба иноземцев: участие иностранцев в обороне южных границ России в 20-х гг. XVII в. // Commentarii de Historia [Электрон. ресурс]. Режим доступа: http://www.main.vsu.ru/~CdH/Articles/07-03.htm.

[3] *Брере Л.* Реляционные базы данных и свободный текст: Contradictio in terminis? // История и компьютер: новые информационные технологии в исторических исследованиях и образовании. Геттинген, 1993. Подобная идея (с некоторыми модификациями) весьма плодотворно используется во многих архивных и библиографических системах, где к структурированному «каталожному» описанию прикрепляется либо машиночитаемая версия полного текста книги (документа), либо их «образы» (в настоящее время чаще всего в формате pdf). В качестве примера см. проект «Открытая русская электронная библиотека» (OREL), размещенный на сайте Российской государственной библиотеки (http://www.rsl.ru/r_frame.asp?http://orel.rsl.ru/) или цифровую библиотеку РНБ (http://www.nlr.ru:8101/e-res/index.html). В России апологетом данного подхода к архивным материалам, относящимся к русскому Средневековью и раннему Новому времени, была главный библиограф РГГУ Е. А. Белоконь. Одним из первых осуществленных ею проектов был проект «Цифровое сохранение и публикация фонда "Древлехранилище" РГАДА».

[4] Собственно, в рамках данной работы был создан «фонд пользования» (в терминах Росархива) тематической коллекции источников по истории иноземцев на русской службе в начале XVII в. Конечно, этот фонд не может быть использован при проведении археографических исследований, однако в случае, когда исследователя интересуют не внешняя критика источника, а его содержание, созданная таблица представляет значительный интерес.

[5] О принципах осуществления подобной работы см.: *Гарскова И. М.* Базы и банки данных в исторических исследованиях. М., 1994

[6] Имеется в виду ситуация, при которой результаты математической обработки существенно меняются в зависимости от изменения набора сведений, включенных в метаисточник.

[7] Описанный прием при видимых недостатках обладает одним скрытым достоинством, т. к. позволяет наглядно продемонстрировать один из главных критериев отбора противоречивой информации — так называемую *частоту упоминания*.