УДК 930.2:004.6
doi 10.17072/2219-3111-2019-3-137-145

МЕЖДУ *DATA* И *CAPTA*: ПРОБЛЕМЫ ДАТАФИКАЦИИ ИСТОРИЧЕСКИХ ИССЛЕДОВАНИЙ[1]

А. Ю. Володин
МГУ имени М.В. Ломоносова,
119192, Москва, Ломоносовский пр., 27, корп. 4
volodin@hist.msu.ru

Рассматривается проблема датафикации исторических исследований: в какой мере существующие и пополняющиеся наборы данных позволяют решать актуальные задачи исторической науки. Датафикация – процесс устойчивого фиксирования массовых наблюдений в разных форматах данных, позволяющий осуществить их качественную и количественную обработку и научный анализ. Понятие «данные» можно определять по-разному: как формат, как объем хранения, как номинальность. Если данные (*data*) часто имеют собственную структуру, которую необходимо приспособить для конкретного исследования, то капта (*capta*) устроена как раз так, чтобы максимально удобно отвечать на вопросы исследования, ради которого она собирается. «Цифровой поворот» все чаще сравнивают с «революцией Гутенберга», в результате которой книгопечатание кардинальным образом изменило представление как о сохранении, так и о распространении и использовании информации. Цифровая эра действительно кажется подобной тем изменениям, которые произошли почти шесть веков назад. Однако мы не найдем в былых эпохах подобия машиночитаемым данным. Но важно учитывать, что данные в исторической науке – не просто зарегистрированные сигналы, они в большинстве случаев являются самостоятельно собранными данными, тщательно, с замыслом и целью решить важную научную проблему, т.е. именно *capta*, которая будет противостоять в исторических исследованиях пусть объективной, но часто излишне общей и сомнительной по содержанию при анализе долгих временных рядов *data*. В современной научной ситуации представляется актуальным сформулировать ясное определение таких понятий, как большие, средние, малые данные. Попытка такого определения в контексте исторической методологии предпринята в данной статье.

Ключевые слова: методология истории, историческая информатика, большие данные, средние данные, малые данные, капта, цифровой поворот, цифровизация, цифровая история.

В любом наборе исходных данных
самая надежная величина,
не требующая никакой проверки,
является ошибочной.
Третий закон Финэйгла

Многообразие окружающих нас электронных ресурсов переносит нас в «эру данных». Данные меняют подход к исследовательским материалам хотя бы потому, что они оказываются недоступными человеку без какого-то специального устройства-посредника (недаром данные часто и сегодня называют машиночитаемыми) [*Маккарти,* 2016; History in the digital age, 2013, p. 1–20]. Такого рода перемены вносят существенные изменения и в исследовательские практики (см. например [*Володин*, 2017]). Правда, влияние новых средств коммуникации на информационную среду замечено было давно. Еще М. Маклюэн выделял период развития медиасреды – «галактику Маркони», которая пришла на смену «галактике Гутенберга» уже больше века назад, с приходом электричества в повседневную коммуникацию [*Маклюэн,* 2005].

Датафикация (или иногда датификация) – процесс устойчивого фиксирования массовых наблюдений в разных форматах данных, позволяющий осуществлять их качественную и количественную обработку и научный анализ. Измерение (т.е. установление соотношения качественных и количественных характеристик) объектов, явлений, процессов реального мира и запись получаемых данных – важная характеристика практически всех обществ письменной истории.

Можно предположить, что датафикация – общий для науки процесс, который может протекать одинаково в разных гуманитарных дисциплинах. Скорее всего, такой взгляд – упрощение. Стоит согласиться с С. Робертсоном в том, что даже при взгляде на, казалось бы, объединенные общей методологической платформой «цифровые гуманитарные науки» нельзя не заметить, что «источники, исследовательские вопросы и подходы, которые они используют в своих проектах, дисциплинарны, равно как дисциплинами определяется выбор цифровых инструментов» [*Робертсон*, 2016, с. 1].

Компьютерная датафикация имеет длительную и насыщенную традицию. «Появление компьютеров повлекло за собой внедрение цифровых устройств для измерения и хранения данных, которые значительно повысили эффективность датафикации, – пишут В. Майер-Шенбергер и К. Кукьер, – а также сделали возможным математический анализ данных для раскрытия их скрытой ценности. Проще говоря, оцифровка стала катализатором датафикации, но никак не ее заменой. Процесс оцифровки (преобразование аналоговой информации в формат, считываемый компьютером) сам по себе не является датафикацией» [*Майер-Шенбергер, Кукьер*, 2014, с. 83].

С практической точки зрения датафикация – это процесс нормализации наблюдений для их систематического анализа. Причем с учетом того, что современные подходы позволяют работать как со структурированными, так и со слабоструктурированными или вовсе неструктурированными данными, уместно говорить не о структурировании данных в соответствии с «нормальной формой», а о гармонизации данных для решения конкретных исследовательских задач [*Володин,* 2016, с. 11]. Гармонизация данных предполагает проведение комплекса мероприятий по повышению степени их согласованности. Сначала процесс гармонизации осуществляется на семантическом уровне, а затем анализируются технологические возможности и ограничения форматов хранения данных в файловой структуре.

Получается, что датафикация является успешной в том случае, когда полученные из исторических источников данные оказываются удобоваримыми для автоматизированного компьютеризированного использования, анализа и управления.

По верному наблюдению М. Таллера, исследователи-гуманитарии сегодня делятся на несколько групп: на исследователей «текста как такового», исследователей-собирателей «фактов» в электронных (иногда весьма обширных) коллекциях, исследователей «нетекстов» (в том числе виртуальных реконструкций), исследователей влияния цифровой среды на гуманитарные науки в целом [*Таллер*, 2012, с. 5–13].

Развитие исторической науки, несмотря на впечатляющие темпы оцифровки исторических документов и успехи оптического распознавания символов, в том числе тотальную оцифровку и создание качественно распознанных электронных архивов, вряд ли достижимо в ближайшей перспективе (хотя все чаще делаются прогнозы этого на 2050 г.). Триада «данные – информация – знания» выходит на первое место в методологических дискуссиях о существовании историко-культурного наследия в разных форматах [*Burke,* 2012; *Корниенко и др.,* 2016; *Нуарэ*, 2017; *Фролов*, 2017; *Эйде*, 2017].

Условно можно выделить несколько типов цифрового представления исторической информации: текст, таблица и изображение, а затем динамические потоковые форматы: аудио и видео – и далее: программные коды (см. подробно о данном наблюдении [*Володин*, 2016]). Особенность «цифрового поворота» можно увидеть в том, что указанные типы на практике чаще всего создают неожиданные сочетания, к примеру: оцифрованная средневековая рукопись первоначально предстает в виде цифрового изображения, затем проводится процедура установления текста, даже если в ней используются компьютерные технологии, например, таблицы базы данных, что в основном согласуется с классическим текстологическим подходом. Таблица придает структуру информации самого разного характера, а связанные таблицы уже могут превратиться в базу данных. Поэтому важным свойством цифровых форматов, которое позволяет увидеть перспективы цифрового подхода, является их многослойность. Необходимость сочетать структурированную и графическую информацию, слабоструктурированную или неструктурированную информацию дает возможность создать основу для мультимедийного понимания современных электронных ресурсов. Многообразие форматов современной передачи информации во многом скрывает от взгляда существенное различие их информационного потенциала, необходимое для актуального осмысления, например, того, что один и тот же файл может сообщить разное количество информации в зависимости от

программы, в которой он открыт. Получается, что исследователь оказывается в прямой зависимости от функционала программного обеспечения. Однако в основе результата процесса оцифровки – данные.

Понятие «данные» можно определять по-разному: (1) как совокупность фактов и идей в формализованном виде, пригодном для передачи и обработки в некотором информационном процессе (в таком случае ключевым свойством данных является их формат); (2) как совокупность ячеек памяти, обладающих определенными свойствами (тогда первостепенным становится вопрос об объеме необходимой памяти для хранения данных и устойчивости носителей хранения); (3) через их номинальность, указывая на то, что в отличие от операций (действий, процессов) данные выражаются подлежащим (с возможными его определениями) [*Володин*, 2015, с. 5]. Тем не менее «данные» как источник исторического исследования меняют привычную источниковедческую перспективу. Мы начинаем смотреть на видовое разнообразие источников не как на объективное, осязаемое различие (наметанный глаз исследователя молниеносно отличает личное письмо от делопроизводственного документа, пусть и хранящегося в одном архивном деле), а как на дополнительное виртуальное свойство, которое может быть как показано (в электронной копии-изображении), так и просто указано в описании к оцифрованному документу.

Уместно вспомнить замечательное рассуждение Ю.М. Лотмана о значении дешифровки для исторической профессии: «Историк обречен иметь дело с текстами. Между событием “как оно есть” и историком стоит текст, и это коренным образом меняет научную ситуацию. Текст всегда кем-то создан и представляет собой происшедшее событие, переведенное на какой-то язык. Одна и та же реальность, кодированная разными способами, даст различные – иногда противоположные – тексты. Извлечение из текста факта, из рассказа о событии – события представляет собой *операцию дешифровки* (выделено мной. – *А.В.*). Таким образом, сознавая это или нет, историк начинает с семиотических манипуляций со своим исходным материалом – текстом» [*Лотман*, 1994, с. 353–354]. При этом сегодня мы смотрим на «текст» как на существующий в нескольких состояниях
объект – рукопись, установленный печатный вариант рукописи и распознанный посимвольно электронный текст в цифровом файле инструментально дают разные (иногда непересекающиеся) возможности для исследователя.

В методологических дискуссиях все чаще отличают *data* (данные) от *capta* (самостоятельно собранные исследователем данные) (см., например [*Chippindale*, 2000, p. 605–612; *Ellis*, 1993, p. 469–73]). Если *данные* часто обладают собственной структурой, которую еще необходимо приспособить для конкретного исследования, то *капта* устроена как раз так, чтобы максимально удобно отвечать на вопросы исследования, ради которого она собирается.

Таким образом, можно заметить и еще одну важную перемену в практике историков: в «эру данных» тексты оказываются в одном ряду с изобразительными, аудиовизуальными и прочими мультимедийными источниками. И, несмотря на сильное сопротивление исконного формата публикации результатов научного исследования в виде монографии, современность требует возможного расширения книжных страниц за счет мультимедийных онлайн-ресурсов [*Cohen, Rosenzweig*, 2005; *Робертсон*, 2016].

«Эра данных» в истории и смежных гуманитарных дисциплинах во многом соответствует международной тенденции изучения «больших данных» (*big data*). Большие массивы данных требуют новых подходов, при этом специализация технологических решений для нужд гуманитарного исследования принципиально необходима, как это было во времена утверждения концепции источнико-ориентированных баз данных. Однако такая перспектива развития часто вступает в противоречие с определением больших данных как потоковых и постоянно пополняемых массивов, а значит, в историческом или гуманитарном исследовании целесообразно использовать прежде всего понятие средних (*medium*) или даже малых (*small*) данных.

В современной историографии все чаще возникают дискуссии о роли данных в актуальных исторических исследованиях, которые так или иначе превращаются в обсуждение роли «больших данных» в современной науке, в том числе в истории. Споры о данных стали важным этапом восприятия историческим сообществом нового этапа развития информационных технологий, а с ними значительно возрос интерес к наработанным подходам и признанным технологиям в рамках таких направлений, как историческая информатика. Такие дебаты оказались новым поводом для оценки

«цифрового поворота» в исторической науке. В качестве примера достаточно указать на активное обсуждение «Исторического манифеста» Д. Гулди и Д. Армитеджа [*Guldi, Armitage*, 2014].

Гулди и Армитедж, определяя причину «кризиса долгосрочного мышления», а вместе с ним и интереса к истории, информационной перегрузкой современных людей, обращаются в главе 4 своей книги к проблеме «больших данных». В частности, они отмечают: «Мы живем в новую эпоху “больших данных” – от расшифровки генома человека до миллиардов слов в официальных отчетах, которые ежегодно производят правительственные учреждения. В социальных и гуманитарных науках обращение историков и социологов к “большим данным” отражает их стремление идти в ногу со временем, использовать открывающиеся возможности для решения старых вопросов и формулирования новых. “Большие данные” стимулируют социальные науки к постановке более масштабных проблем. В истории это прежде всего события мирового масштаба и длительная институциональная динамика. В проектах, посвященных долгосрочной истории изменений климата, последствиям работорговли или разнообразию форм права собственности на Западе, использование вычислительных методов позволяет исследователям открывать новые аспекты работы с данными и связывать исторические проблемы с современными» [*Guldi, Armitage*, 2014, p. 88].

Стоит заметить, что данное поветрие в мировой науке стимулирует вовлечение в дискуссию самого широкого круга ученых, часто весьма условно представляющих роль данных в целом и баз данных в частности, фактически уже многие десятилетия использующихся в исторической науке [*Гарскова*, 2018; *Бородкин*, 2016].

Бесспорно, технология баз данных стала базовым методом информатизации исторической науки (причем они даже составляют технологическую основу столь популярных сегодня визуальных расширений исторического инструментария – ГИС- и 3D-реконструкций) [*Гарскова*, 1994; *Бородкин*, 2016, с. 258–276]. Датафикация исторического знания требует переведения определения масштабов данных (столь привычных для имевших место дискуссий) из метафорического в инструментальное проблемное поле, что позволит избежать ошибок и найти пути практического применения данных разных масштабов в современных исторических исследованиях.

Так, понятие «большие данные» (*big data*), очевидно, в качестве определяющего для исторических источников понадобится будущим исследователям современной эпохи (конечно, при условии сохранения имеющихся сегодня комплексов больших данных). С практической точки зрения любые исторические данные, даже исключительно больших объемов, не будут относиться к машинопорождаемым неструктурированным потоковым данным значительного многообразия и огромного объема (показательной является монография [*Manning*, 2013]). Однако инструменты работы с такими данными, вроде фреймворка Hadoop или подхода noSQL, могут оказаться применимы к слабоструктурированным историческим данным.

В качестве примера можно обратиться к монографии П. Мэннинга [*Manning*, 2013], в которой обобщается англоязычная литература о применении «больших данных» в исторических исследованиях начиная с первых опытов создания баз данных. Примером удачного опыта сотрудничества в области накопления и анализа исторической информации можно считать проект CHIA (Collaborative for Historical Information and Analysis)[2]. Мэннинг определяет несколько последовательных целей, которые стоят перед исследователями «больших данных»: 1) сбор и описание исторических данных; 2) создание детального исторического архива данных; 3) анализ и визуализация исторических данных во всемирном масштабе. При этом одной из принципиальных задач может стать глобальное сравнение исторической динамики. Работа с «большими данными», очевидно, требует сотрудничества различных специалистов и состоит из нескольких этапов: 1) получение данных (в том числе с помощью популярного сетевого краудсорсинга, под которым часто понимают привлечение множества волонтеров к трудозатратной деятельности, которую можно выполнить онлайн (см., например [*Куликов*, 2016]); 2) документирование данных (в том числе метаописание, курирование и гармонизация); 3) обработка данных (включая агрегирование и «добычу» данных); 4) анализ данных (позволяющий создавать модели и теории); 5) визуализация данных (в частности, создание интерфейса для запросов и выдачи результатов обработки данных). Именно такая информационная архитектура использована в указанном проекте CHIA.

Понятие «средние данные» (*medium data*) все чаще применяется в научной литературе для описания крупных коллекций данных, которые не претендуют на поток постоянных пополнений в духе «больших» данных. Тем не менее использование данных средних объемов может оказаться

наиболее удобным при апробации новых подходов, так как они позволяют строго контролировать имеющиеся переменные и наблюдения и создают возможности для построения взаимосвязей – графиков, карт (например, на платформе *Tableau* или в среде *RStudio*), сетевого анализа (например, *Ucinet* и *NetDraw*), а также автоматической семантической разметки (например, с помощью *MaxQDA* или *n-gram*). Недостаток же «средних данных» состоит в том, что обычно они отбираются из больших коллекций по одному или нескольким техническим параметрам и решение о репрезентативности их принимает исследователь [*Орлова*, 2016].

Понятие «малые данные» (*small data*) все чаще употребляется в качестве противопоставления: «малые данные», собранные исследователем самостоятельно (как раз речь идет о капте), отличаются от «больших» и «средних» данных, полученных (скачанных/загруженных) из различных депозитариев научной информации. Обращение к «малым данным» позволяет проявить исследовательские компетенции при построении инфологической и даталогической моделей. «Малые данные» полностью контролируются исследователем, при этом обычно подробно обосновываются репрезентативность и целостность самостоятельно собранных данных, а для нужд вторичного использования осуществляется подробное документирование исторических источников таких коллекций данных.

На примере данных, применяемых в социально-экономической истории, можно проследить эволюцию представления данных – от «малых данных» (конкретно-исторических таблиц) до «средних данных» (взаимосвязанных и подробно описанных и размеченных SQL-таблиц с возможностью как выгрузки данных, так и онлайн-анализа) – и выделить несколько поколений репозитариев исторических данных (более подробную классификацию см. [*Володин*, 2019]):

• 1960–1970 гг:: HRAF, ICPSR, NBER, OECD, WB[3];

• 1980–1990 гг.: UNPOP, IPUMS, GENH, IROWS[4];

• 2000–2010 гг.: CLIO-INFRA, GapMinder, «Динамика…», РИСтат[5].

Проектов, в которых используются полноценные «большие данные» для конкретно-исторических исследований, пока немного. Однако объемы оцифрованных и распознанных данных существенно увеличиваются. И появляются работы, основанные не на тысячах или сотнях наблюдений, а на миллионах записей разнородных данных (часто подобные исследования встречаются в демографической истории (см., например [Historical Methods, 2011; *Schürer*, 2007]). Недаром в цифровых исследованиях последних лет (особенно в рамках «цифровой гуманитаристики», или «digital humanities») закрепилось различение «близкого» и «дальнего» чтения [*Моретти*, 2016б, с. 6]. Разница в том, что при «близком» чтении (*close reading*) приходится полагаться на себя и свои способности в поиске смыслов в прочитанном, а при «дальнем» чтении (*distant reading*) большое внимание уделяется алгоритму поиска, выявления и сопоставления нужных фрагментов в значительном корпусе источников (часто существенно превышающем физические возможности их прочесть) [*Моретти*, 2016а, с. 10].

Таким образом, максимально разработанные в историографической традиции «малые данные» имеют ясную перспективу в исследовательских практиках: понятно, какие вопросы можно задать и как на них ответить с помощью таких данных. Использование же «средних данных» требует скорейшей инструментализации, так как от этого во многом зависит, в какой степени могут состояться исторические исследования с уже оцифрованными обширными коллекциями исторических источников (например, периодики). Вероятно, можно взять на вооружение принципы работы с разными типами данных (например, 5V для «больших данных»: volume/объем, velocity/скорость, variety/разнообразие, variability/разнообразие, value/ценность), чтобы перейти к конкретному инструментальному пониманию существующего разнообразия данных и подходов к их анализу, необходимых в современных исторических исследованиях.

Очевидно, что выбор эффективных средств анализа и *data*, и *capta* зависит от нужд конкретно-исторических исследований и возможностей конкретных исследователей истории. Причем суть «цифрового поворота» в исторической науке имеет смысл видеть в новом понимании информационного моделирования, особого подхода к реконструкции прошлого на основе данных, что существенно влияет и на современные исследовательские практики историков. Новой чертой цифровой эпохи становится подход к изученным материалам как данным, требующим не только компьютерной обработки и визуализации, но и экспертной интерпретации с учетом возможностей современных методов обработки данных.

«Цифровой поворот» все чаще сравнивают с «революцией Гутенберга», в результате которой книгопечатание кардинальным образом изменило представление как о сохранении, так и о распространении и использовании информации. Цифровая эра, действительно, кажется подобной тем изменениям, которые произошли почти шесть веков назад. Однако мы не найдем прямого подобия машиночитаемым данным в прошлом, но новые технологии необходимо встраивать в долгосрочную историческую традицию источниковедения. И важно учитывать, что данные в исторической науке – не просто зарегистрированные сигналы, они в большинстве случаев собраны с целью решить важную научную проблему, т.е. *capta*, которая будет в исторических исследованиях противопоставляться пусть объективной, но часто излишне общей и сомнительной по содержанию при анализе долгих временных рядов *data*.

Примечания

[1] Публикация подготовлена при финансовой поддержке РФФИ, грант № 17-01-50134.

[2] CHIA. URL: http://www.chia.pitt.edu. В качестве примеров онлайн-проектов по сбору исторической информации также часто указываются: CLIO-INFRA – clio-infra.eu (Амстердам), CODESRIA – codesria.org (Дакар) и CLACSO – clacso.org.ar (Буэнос-Айрес).

[3] Human Relations Area Files. URL: hraf.yale.edu; The Interuniversity Consortium for Political and Social Research. URL: icpsr.umich.edu; The National Bureau of Economic Research. URL: nber.org; Organisation for Economic Co-operation and Development. URL: oecd.org; World Bank URL: worldbank.org.

[4] United Nations Population Division. URL: www.un.org/esa/population; Integrated Public Use Microdata Series. URL: ipums.org; Global Economic History Network. URL: lse.ac.uk/economicHistory/Research/GEHN; Institute for Research on World–Systems. URL: irows.ucr.edu; Maddison Project. URL: ggdc.net/Maddison.

[5] CLIO-INFRA. URL: www.clio-infra.eu; GapMinder. URL: www.gapminder.org; Динамика экономического и социального развития России в XIX – начале XX в. URL: http://hist.msu.ru/Dynamics; Электронный архив Российской исторической статистики (РИСтат). URL: http://ristat.org.

Библиографический список

Бородкин Л.И. Моделирование исторических процессов: от реконструкции к анализу альтернатив. СПб.: Алетейя, 2016. 310 с.

Володин А.Ю. «Цифровая история»: ремесло историка в цифровую эпоху // Электрон. науч.-образов. журнал «История». 2015. Т. 6. Вып. 8 (41). URL: https://history.jes.su/s207987840001228-9-1/ (дата обращения: 03.06.2019).

Володин А.Ю. Цифровой поворот в исторической науке: вероятное и неочевидное // Электрон. науч.-образов. журнал «История». 2016. Т. 7. Вып. 7 (51). URL: https://history.jes.su/s207987840001647-0-1/ (дата обращения: 03.06.2019).

Володин А.Ю. Цифровые практики ученых-гуманитариев: результаты онлайн-исследования // Электрон. науч.-образов. журнал «История». 2017. Т. 8. Вып. 7 (61). URL: https://history.jes.su/s207987840001967-2-1 (дата обращения: 03.06.2019).

Володин А.Ю. Онлайн-репозитории статистических данных по социально-экономической истории: возможности и перспективы // Отечественные архивы. 2019. № 3. С. 34–42.

Гарскова И.М. Базы и банки данных в исторических исследованиях. М., 1994. 215 с.

Гарскова И.М. Историческая информатика. Эволюция междисциплинарного направления. СПб.: Алетейя, 2018. 408 с.

Корниенко С.И., Гагарина Д.А., Поврозник Н.Г. Информационные системы в цифровой среде исторической науки // Электрон. науч.-образов. журнал «История». 2016. Т. 7. Вып. 7 (51). URL: https://history.jes.su/s207987840001638-0-1/ (дата обращения: 03.06.2019).

Куликов В.А. Краудсорсинг в сохранении и изучении культурного наследия // Электрон. науч.-образов. журнал «История». 2016. Т. 7. Вып. 7 (51). URL: https://history.jes.su/s207987840001670-6-1/ (дата обращения: 03.06.2019).

Лотман Ю.М. Изъявление Господне или азартная игра? (Закономерное и случайное в историческом процессе) // Ю.М. Лотман и тартуско-московская семиотическая школа. М.: Б. и., 1994. С. 353–363.

Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим. М.: Б. и., 2014. 240 с.

Маккарти У. Специальные эффекты: инструменты есть, а где результаты? // Электрон. науч.-образов. журнал «История». 2016. Т. 7. Вып. 7 (51). URL: https://history.jes.su/s207987840001637-9-1/ (дата обращения: 03.06.2019).
Маклюэн М. Галактика Гутенберга. Становление человека печатающего. М.: Б. и., 2005. 496 с.
Моретти Ф. Дальнее чтение. М.: Дело, 2016. 352 с.
Моретти Ф. Масштаб, смысл, паттерн, форма: концептуальные задачи, стоящие перед количественным литературоведением // Электрон. науч.-образов. журнал «История». 2016б. Т. 7. Вып. 7 (51). URL: https://history.jes.su/s207987840001646-9-1/ (дата обращения: 03.06.2019).
Нуарэ С. Цифровая история: история и память, доступные каждому // Электрон. науч.-образов. журнал «История». 2017. Т. 8. Вып. 7 (61). URL: https://history.jes.su/s207987840001917-7-1/ (дата обращения: 03.06.2019).
Орлова Г.А. Е-Оксюморон: дигитальное как качественное // Электрон. науч.-образов. журнал «История». 2016. Т. 7. Вып. 7 (51). URL: https://history.jes.su/s207987840001644-7-1/ (дата обращения: 03.06.2019).
Робертсон С. Различия между цифровыми гуманитарными науками и цифровой историей // Электрон. науч.-образов. журнал «История». 2016. Т. 7. Вып. 7 (51). URL: https://history.jes.su/s207987840001648-1-1/ (дата обращения: 03.06.2019).
Таллер М. Дискуссии вокруг Digital Humanities // Историческая информатика. 2012. № 1. С. 5–13.
Фролов А.А. «Цифровой поворот» и крупномасштабные исторические карты в отечественной историографии // Электрон. науч.-образов. журнал «История». 2017. Т. 8. Вып. 7 (61). URL: https://history.jes.su/s207987840001957-1-1/ (дата обращения: 03.06.2019).
Эйде О. Вербальное выражение географической информации // Электрон. науч.-образов. журнал «История». 2017. Т. 8. Вып. 7 (61). URL: https://history.jes.su/s207987840001948-1-1/ (дата обращения: 03.06.2019).
Burke P. A Social History of Knowledge II: From the Encyclopaedia to Wikipedia. Cambridge: Polity Press, 2012. 248 p.
Chippindale C. Capta and Data: On the True Nature of Archaeological Information // American Antiquity. 2000. Vol. 65, No. 4, Oct. P. 605–612.
Cohen D.J., Rosenzweig R. Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web. University of Pennsylvania Press, 2005. 328 p.
Ellis D. Modeling the Information-Seeking Patterns of Academic Researchers: A Grounded Theory Approach // The Library Quarterly. 1993. № 63 (4). P. 469–486.
Guldi J., Armitage D. The History Manifesto. Cambridge University Press, 2014. 175 p.
Historical Methods: A Journal of Quantitative and Interdisciplinary History. 2011. Vol. 44, Issue 1: Big Data: Large-Scale Historical Infrastructure from the Minnesota Population Center.
History in the digital age / Ed. T. Weller. London; New York: Routledge, 2013. 226 p.
Manning P. Big data in history. Palgrave, 2013. 129 p.
Schürer K. Creating a Nationally Representative Individual and Household Sample for Great Britain, 1851 to 1901 – The Victorian Panel Study (VPS) // Historical Social Research. 2007. 32. P. 211–331.

Дата поступления рукописи в редакцию 18.06.2019

BETWEEN ‘DATA’ AND ‘CAPTA’: THE PROBLEM OF DATAFICATION IN HISTORICAL RESEARCH

A. Y. Volodin

Moscow State University, Lomonosovskiy ave., 27-4, 119192, Moscow, Russia
volodin@hist.msu.ru

The article deals with the problem of datafication in historical research. The author discusses to what extent existing and growing data sets allow historians to solve actual problems of history. Datafication is a process of stable recording of mass observations in different data formats with the possibility of their qualitative and quantitative processing and scientific analysis. Data can be understood and defined in different ways: as a format, as a storage volume, or as a nominal value. If the data often has its own structure, which still needs to be adapted for a particular

study, the capta is arranged in the most convenient way to answer the questions of the concrete research for which it is collected. The “digital turn” is being increasingly compared with the “Gutenberg revolution", when book printing radically changed the way we preserve, disseminate and use information. The digital age does something similar to the changes that took place almost six centuries ago. However, in past eras, we cannot find direct similarities to machine-readable data. It is important to take into account that the data in historical research are not just registered signals, but in most cases they are carefully, intentionally and purposefully collected bits of information to solve an important problem, that is the very capta, which will resist to the data which is objective, but too general and in some places dubious in the analysis of long time series. In the current scientific situation, it seems relevant to formulate a clear definition of such concepts as large, medium, or small data.

Key words: methodology of history, historical information science, big data, medium data, small data, capta, digital turn, digitalization, digital history.

References

Borodkin, L.I. (2016), *Modelirovanie istoricheskikh protsessov: ot rekonstruktsii k analizu al'ternativ* [Modeling of historical processes: from reconstruction to analysis of alternatives], Aleteiya, St. Petersburg, Russia, 310 s.

Burke, P. (2012), A *Social History of Knowledge II: From the Encyclopaedia to Wikipedia*, Polity Press, Cambridge, UK, 248 p.

Chippindale, C. (2000), “Capta and Data: On the True Nature of Archaeological Information”, *American Antiquity*, Vol. 65, No. 4, Oct., pp. 605–612.

Cohen, D.J. & R. Rosenzweig (2005), *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*, University of Pennsylvania Press, Pennsylvania, USA, 328 p.

Eide, Ø. (2017), “Verbal Expressions of Geographical Information”, *Elektronnyy zhurnal “Istoriya”*, Vol. 8, issue 7 (61), available at: https://history.jes.su/s207987840001948-1-1/ (accessed 03.06.2019).

Ellis, D. (1993), “Modeling the Information-Seeking Patterns of Academic Researchers: A Grounded Theory Approach”, *The Library Quarterly*, № 63 (4), pp. 469–486.

Frolov, A. A. (2017) “Digital turn” and large-scale historical maps in Russian historiography”, *Elektronnyy zhurnal “Istoriya”*, Vol. 8, issue 7 (61), available at: https://history.jes.su/s207987840001957-1-1/ (accessed 03.06.2019).

Garskova, I.M. (1994), *Bazy i banki dannykh v istoricheskikh issledovaniyakh* [Databases and data banks in historical research], Konrad Pachnicke Max-Planck-Institut fur Geschichte, Göttingen, Germany, 215 p.

Garskova, I.M. (2018) *Istoricheskaya informatika. Evolyutsiya mezhdistsiplinarnogo napravleniya*. [Historical Information Science. Evolution of interdisciplinary field of research] SPb.: Aleteiya, 408 p.

Guldi, J. & D. Armitage (2014), *The History Manifesto*, Cambridge University Press, Cambridge, Great Britain, 175 p.

Historical Methods: A Journal of Quantitative and Interdisciplinary History (2011). Vol. 44, Issue 1: Big Data: Large-Scale Historical Infrastructure from the Minnesota Population Center.

Weller, T. (2013), *History in the digital age*, Routledge, London; New York, UK; USA, 226 p.

Kornienko, S.I., Gagarina, D.A. & N.G. Povroznik (2016), “Information systems in the digital environment of historical research”, *Elektronnyy zhurnal “Istoriya”*, Vol. 7, issue 7 (51), available at: https://history.jes.su/s207987840001638-0-1/ (accessed 03.06.2019).

Kulikov, V.A. (2016), “Crowdsourcing in the preservation and study of cultural heritage”, *Elektronnyy zhurnal “Istoriya”,* Vol. 7, issue 7 (51), available at: https://history.jes.su/s207987840001670-6-1/ (accessed 03.06.2019).

Lotman, Yu.M. (1994), “Expression of the Lord or just gambling? (Natural and accidental in the historical process)”, in *Yu.M. Lotman i tartusko-moskovskaya semioticheskaya shkola* [Yu.M. Lotman and the Tartu-Moscow semiotic school], Gnozis, Moscow, Russia, pp. 353–363.

Manning, P. (2013), *Big data in history*, Palgrave, ГОРОД, СТРАНА, 129 p.

Mayer-Shenberger, V. & K. Kuk'er (2014), *Bol'shie dannye. Revolyutsiya, kotoraya izmenit to, kak my zhivem, rabotaem i myslim* [Big Data: The Essential Guide to Work, Life and Learning in the Age of Insight], Mann, Ivanov, Farber, Moscow, Russia, 240 p.

Makkarti, W. (2016), “Special Effects or, The Tooling is Here. Where are the Results?”, *Elektronnyy zhurnal “Istoriya”,* T. 7, vyp. 7 (51), available at: https://history.jes.su/s207987840001637-9-1/ (accessed 03.06.2019).

Makluen, M. (2005), *Galaktika Gutenberga. Stanovlenie cheloveka pechatayushchego* [The Gutenberg Galaxy: The Making of Typographic Man], Akademicheskiy Proekt, Moscow, Russia, 496 p.

Moretti, F. (2016), *Dal'nee chtenie* [Distant reading], w.p., Moscow, Russia, 352 p.

Moretti, F. (2016b), “Scale, Meaning, Pattern, Form: Conceptual Challenges for Quantitative Literary Studies”, *Elektronnyy zhurnal “Istoriya”*, Vol. 7, issue 7 (51), available at: https://history.jes.su/s207987840001646-9-1/ (accessed03.06.2019).

Noiret, S. (2017), “Digital History: History and Memory — Accessible for Everyone”, *Elektronnyy zhurnal “Istoriya”*, Vol. 8, issue 7 (61), available at: https://history.jes.su/s207987840001917-7-1/ (accessed 03.06.2019).
Orlova, G. A. (2016), “-oxymoron: Digital as Qualitative”, *Elektronnyy zhurnal “Istoriya”*, Vol. 7, issue 7 (51), available at: https://history.jes.su/s207987840001644-7-1/ (accessed 03.06.2019).
Robertson, S. (2016), “The Differences between Digital Humanities and Digital History”, *Elektronnyy zhurnal “Istoriya”*, Vol. 7, issue 7 (51), available at: https://history.jes.su/s207987840001648-1-1/ (accessed 03.06.2019).
Taller, M. (2012), “Debate around Digital Humanities”, *Istoricheskaya informatika*, № 1. p. 5–13.
Volodin, A.Yu. (2015), ““Digital history”: the craft of a historian in the digital age”, *Elektronnyy zhurnal “Istoriya”*, Vol. 6, issue 8 (41), available at: https://history.jes.su/s207987840001228-9-1/ (accessed 03.06.2019).
Volodin, A.Yu. (2016), “Digital turn in historical research: obvious and non-obvious”, *Elektronnyy zhurnal “Istoriya”*, Vol. 7, issue 7 (51), available at: https://history.jes.su/s207987840001647-0-1/ (accessed 03.06.2019).
Volodin, A.Yu. (2017), “Digital practices of humanitarians: results of online research”, *Elektronnyy zhurnal “Istoriya”*, Vol. 8, issue 7 (61), available at: https://history.jes.su/s207987840001967-2-1 (accessed 03.06.2019).
Volodin, A.Yu. (2019), “Online repositories of statistical data on socio-economic history: opportunities and prospects”, *Otechestvennye arkhivy*, № 3, pp. 34–42.