DOI 10.37882/2223-2982.2024.7-2.05

ЦИФРОВЫЕ ГУМАНИТАРНЫЕ ИССЛЕДОВАНИЯ В КОНТЕКСТЕ ДАТАИЗМА

DIGITAL HUMANITIES IN THE CONTEXT OF DATAISM

A. Volodin

Summary: The current state of art and key trends of an interdisciplinary approach known as the digital humanities is examined in the article. The definitions of the digital humanities, the latest publications on this issue are critically reviewed, special attention is paid to the current tasks of interdisciplinary digital research using the examples of success in the digitalization of historical and cultural heritage.

Keywords: digital humanities, data science, digitization, digital turn, online resources.

Володин Андрей Юрьевич
кандидат исторических наук, доцент, Московский государственный университет имени М.В. Ломоносова; ведущий научный сотрудник, Сибирский федеральный университет (Красноярск)
volodin@hist.msu.ru

Аннотация: В статье кратко рассмотрены современное состояние и тенденции развития междисциплинарного научного направления – цифровых гуманитарных исследований или digital humanities. Критически представлены определения цифровых гуманитарных наук, новейшая литература по данной проблематике, особое внимание уделяется вопросам текущих задач междисциплинарных исследований с использованием успехов цифровизации историко-культурного наследия.

Ключевые слова: цифровые гуманитарные науки, наука о данных, оцифровка, цифровой поворот, онлайн-ресурсы.

Digital Humanities – цифровые гуманитарные исследования или цифровая гуманитаристика – актуальное направление, в последние два десятилетия активно завоевывающее место среди гуманитарных междисциплинарных компьютеризированных исследованиях [7, 8]. Компьютеризация началась в гуманитарных науках не сегодня, ведь к помощи компьютерной техники в гуманитарных исследованиях обратились с появлением больших вычислительных машин в 1970-е годы. Но цифровая эпоха в гуманитарные науки пришла после микрокомпьютерной революции с развитием вычислительных мощностей и персонализации компьютерных систем, позволяющих не только создавать сложные виртуальные реконструкции, но и представлять их в электронной среде с помощью средств Всемирной паутины.

Можно уверенно сказать, что цифровой поворот в гуманитарных исследованиях состоялся. И сегодня лишь отшельник может не соприкасаться с цифровой цивилизацией. Исследователи-гуманитарии сегодня используют спонтанную и систематическую, выборочную и сплошную оцифровку документов, памятников и объектов историко-культурного наследия. Оцифровка или работа с оцифрованными коллекциями стала одной из важных ежедневных практик ремесла гуманитария. В этой связи встает широкий спектр вопросов, в чем преимущества и недостатки наступления цифровой эры в гуманитарных исследованиях – именно эти вопросы оказываются во главе угла в весьма обширной литературе, посвященной проблемам определения, самоопределения и развития междисциплинарного направления цифровых гуманитарных наук [7, 10, 13].

Следует обратить особое внимание на коллективную монографию «Цифровые гуманитарные исследования» [7], в которой впервые на русском языке комплексно рассмотрено это актуальное междисциплинарное направление. В книге приведены примеры (само)определения направления, дан их обзор. «Цифровой поворот» в гуманитарных исследованиях и масштабные проекты оцифровки историко-культурного наследия описаны в контексте датафикации и вызовов больших данных и машинного обучения. Особое внимание уделено современным подходам к компьютерному анализу текстов и культуромике, направлению исследований культуры и языка с помощью больших текстовых данных. Представлена широкая палитра цифровых подходов, призванных находить решения насущных гуманитарных исследовательских задач: от базы данных к сетевому анализу, от геоинформационных систем к виртуальным реконструкциям и дополненной реальности. Происходящие процессы рассмотрены в связи со становлением сложной и противоречивой информационной инфраструктуры цифровых гуманитарных исследований.

Действительно, многие гуманитарные дисциплины весьма успешно включились в процесс использования цифровых технологий для решения научных задач, лидерами в этом стали филология и история. Филологи значительно продвинулись в компьютеризированном изучении текстов, создании лингвистических корпусов, автоматизации процедур текстологического анализа. Историки сосредоточились на изучении оцифрованных исторических источников, представлении исторических сведений в формате баз данных, оцифровке и электронной публикации свидетельств прошлого.

При этом, более всего актуальным становится переход от измерительных технологий к реконструкциям, связанным с быстрым развитием средств компьютерной визуализации, распространением сетевых технологий, бурным развитием больших языковых моделей. Данный переход можно условно датировать 2004 г., когда постепенно стало заметно терминологическое изменение: от исторического или гуманитарного компьютинга (humanities computing, history and computing) терминология начала переходить к цифровым гуманитарным наукам [8]. Перемена названия означала постепенное изменение статуса – от технической поддержки к интеллектуальному прорыву со своими профессиональными практиками, научными стандартами и теоретическими построениями. Во многом переход от «измерительных» возможностей компьютерных технологий к реконструкционным и презентационным связан с освоением интернет-технологий в разных гуманитарных областях, которые в последнее время усилились успехами векторизации текстовых файлов и скоростным развитием технологий на базе генеративных предобученных трансформеров.

Новым взглядом на современные изменения в гуманитарном познании окружающего мира является ***датаизм***. «Датаизм провозглашает, что Вселенная состоит из потоков данных и что ценность всякого явления или сущности определяется их вкладом в обработку данных» – так Ю.Н. Харари описывает современную эпоху увлеченности «большими данными» в книге «Homo Deus: краткая история будущего» [6, с. 430]. «Датаизм разрушает барьер между животными и машинами и предрекает, – продолжает он, – что электронные алгоритмы в конце концов расшифруют и превзойдут биохимические алгоритмы». Но смысл датаизма как подхода к современной цифровой реальности не столько в противопоставлении человека и машины, а в изменении ключевого процесса перевода зафиксированных сигналов реальности в мудрость. Классическая схема информационной иерархии состоит в том, что сигналы складываются в данные, из данных человек создает информацию, информация уже преобразовывается в знания, а знания позволяют достичь мудрости. «Датаизм переворачивает традиционную пирамиду обучения. До недавних пор на данные смотрели как на первое звено в длинной цепочке интеллектуальной деятельности. Человеку надо было превращать данные в информацию, информацию в знания, а знания в мудрость. Но датаисты считают, что люди больше не в состоянии справляться с огромными потоками данных, поэтому не могут превращать данные в информацию и уж тем более в знания или мудрость. Поэтому обработка данных должна быть доверена электронным алгоритмам, намного более мощным, чем человеческий мозг. На практике это означает, что датаисты скептически относятся к человеческим знаниям и мудрости и предпочитают полагаться на большие данные и компьютерные алгоритмы» [6, с. 430].

В контексте датаизма существенно меняется фокус внимания исследователей в дискуссиях об особенностях формализации сведений гуманитарных источников в моделях данных. Сегодня можно встретить немало противоречий в рассуждениях о принципах применения информационных технологий в гуманитарных исследованиях, связанных со сложностями взаимодействия подходов информатики, гуманитарных и социальных наук, все еще встречаются противопоставления качественных и количественных подходов, несмотря на широкое распространение «смешанных методов» и подходов цифровых гуманитарных исследований (Digital Humanities), а также отмечаются различия задач информационных технологий как отрасли и научного сообщества. Тем не менее, данные становятся своего рода общим знаменателем таких дискуссий, потому что в современных исследованиях приходится опираться именно на данные. Данные в гуманитарных исследованиях позволяют абстрагироваться от исходных источников и собрать систематические формализованные наблюдения. В таком случае данные уже рассматриваются в технологическом контексте как формат (с возможностями и ограничениями каждого конкретного формата), занимаемая память (с ограничениями по объему и устойчивости хранения) и пассивность (противопоставляемая активности кода компьютерных программ). Особое внимание в таком случае необходимо уделить таким свойствам научных данных, как объем, разнообразие форматов, скорость накопления, изменчивость, достоверность, визуализируемость и ценность. Гуманитарные данные в семиотическом смысле выполняют три ключевые функции: называют свойства предметов реального мира (номинация), связывают названные свойства друг с другом (предикация), располагают названное в пространстве и времени (локация). Таким образом, с методологической точки зрения, в гуманитарном исследовании данные – это метод фиксирования абстрактного наблюдения, когда из разнообразных источников последовательно и строго формализованно собираются систематические данные [см. подробнее: 1].

Данные – это подход к регистрированию явлений действительности, претендующий на абсолютную формализацию не только хранения зарегистрированной информации, но и ее познания. Регистрируемость явлений и событий – сложная проблема в гуманитарном познании. Многообразие существующих и создающихся электронных ресурсов переносит нас в «эру данных». Данные меняют подход к исследовательским материалам хотя бы потому, что они оказываются недоступными человеку без специального устройства-посредника (недаром данные часто и сегодня называют машиночитаемыми) [3]. Такого рода перемены вносят существенные изменения и в исследовательские практики. Правда,

влияние новых средств коммуникации на информационную среду замечено было давно. Ещё М. Маклюэн выделял период развития медиасреды – «галактику Маркони», которая пришла на смену «галактике Гутенберга» уже больше века назад, с приходом электричества в повседневную коммуникацию [4].

Датафикация – процесс устойчивого фиксирования массовых наблюдений в разных форматах данных, позволяющий осуществлять их качественную и количественную обработку и научный анализ. Измерение (то есть установление соотношения качественных и количественных характеристик) объектов, явлений, процессов реального мира и запись получаемых данных – важная характеристика практически всех обществ письменной истории. По сути, датафикация – общий для науки процесс, который может протекать одинаково в разных гуманитарных дисциплинах. Хотя каждый специалист будет настаивать на принципиальных отличиях данных в собственной научной области. Как например С. Робертсон отмечает, что несмотря на общую методологическую платформу под названием «цифровые гуманитарные науки», «источники, исследовательские вопросы и подходы, которые они используют в своих проектах, дисциплинарны, равно как дисциплинами определяется выбор цифровых инструментов» [5].

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» [2, с. 83].

С научной 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 слабо структурированными или вовсе неструктурированными данными, уместно говорить не о структурировании данных в соответствии с «нормальной формой», а о гармонизации данных для решения конкретных исследовательских задач. 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 оказывается успешной в том случае, когда полученные из объектов исследования данные оказываются удобоваримыми для автоматизированного компьютеризированного использования, анализа и управления.

Данные – это абстракции сущностей реального мира (человека, объекта или события). Данными могут стать любые зафиксированные сигналы. Данные состоят из свойств, или переменных, или атрибутов, как бы мы не привыкли называть конкретный вид абстракции. Каждый объект обычно описывается рядом атрибутов. Простой пример, хорошо известный любому гуманитарию – книга, которая может иметь следующие свойства: автор, название, издательство, место и год издания, количество страниц, ISBN, цена, жанр, тема и т.д. Важным навыком цифрового гуманитария является умение рассматривать текст как данные [12].

В гуманитарных науках всё чаще начинает использоваться понятие к***а***пта [11]. Что такое капта? Образно говоря, это исследовательский «улов». Это те данные, которые историк собрал в архиве, лингвист, например, в поле, а философ на ментальной карте. Иногда такие данные называют «естественной выборкой», понимая под ней те оставшиеся источники, документы, артефакты прошлого, которые мы можем анализировать. Фактически это собранные доступные данные. Для понимания «капты» может помочь образ археологического раскопа. То, что найдено в раскопе в этом году является последней по близости к настоящему находкой, но лишь очередной на поступательном пути науки. В следующем году будут новые находки, но анализировать и интерпретировать мы можем только то, что есть у нас в руках сегодня. Да, будущие открытия дополнят наши знания, но исследуем мы то, что есть.

Для исследовательских задач работы с данными требуется внимательное отношение к семантическому моделированию данных. Такой подход строится на понимании смысла этих данных. Причем различные исследования могут вкладывать в одни и те же данные разные смыслы, видеть разные связи, проверять различные гипотезы. Недаром сегодня стала популярной концепция FAIR по отношению к исследовательским данным. Акроним FAIR описывает так называемые честные данные, которые отвечают четырем требованиям: в них можно осуществлять поиск, они должны быть доступны, такие данные должны быть интероперабельны, то есть совместимы с разными программами и операционными системами, и наконец, они должны позволять многоразовое использование.

Почему это действительно важно? В идеальном мире каждое новое исследование должно привносить в копилку науки новые наборы данных, базы данных, оциф-

рованные документы и артефакты. И все эти материалы, словно кусочки смальты, шаг за шагом должны создавать или воссоздавать огромное мозаичное панно истории и культуры. Но в реальности часто оказывается, что наборы данных перестают быть доступными, совместимыми или воспроизводимыми, как только конкретное исследование заканчивается, завершается проект, или пропадает интерес у исследователя.

Для целей сохранения добытых данных создаются репозитории — специальные хранилища. В каждой предметной области создаются такие специализированные информационные системы. А нужно не забыть, что информационные системы — это не только серверы и программы, но и люди, их навыки, время и отзывы. В каких-то областях репозитории успешнее — например, корпусы в лингвистике или геоинформационные системы в истории, так как накопление данных общепризнано в сообществе. В других случаях сбор данных происходит медленно. Но суть требований к современным данным это не меняет. Если Вы рассчитываете не только получить результат, но и сохранить данные для будущих поколений исследователей, необходимо подробно задокументировать свои находки и подыскать для них репозиторий с хорошей репутацией. Остается надеяться, что цифровые гуманитарные исследования смогут стать опорой для глобальных возможностей в развитии исследований, а не очередным поводом для «цифрового разрыва», ведь всё очевиднее становится, что цифровой инструментарий в условиях необходимости индустриального производства цифровых научных результатов становится весьма дорогим удовольствием [9].

За последние годы уже сложился определенный «канон» компьютерных методов, который необходимо иметь в виду обращаясь к машиночитаемым данным или рассчитывая увидеть смысл в оцифрованных коллекциях, превышающих возможности нашего физического восприятия. К такому «канону» можно отнести базы данных, компьютерный анализ текста, геоинформационный анализ, сетевой анализ данных, компьютерное моделирование [см. 7, 10, 13]. При этом в большинстве случаев эти методы — это в исконном значении путь исследования, который каждый гуманитарий проходит по-своему, собирая уникальное соотношение нужных подходов и умений, которые существенно обогащают и так присущие исследователям навыки вдумчивого чтения, пристального наблюдения и интуитивной классификации. Чем более сложным будет инструментарий исследователя, тем сильнее он будет себе казаться. Но не стоит обольщаться, полагая, что методы заменят знания. Важно понимать, что цифровые методы и данные ни в коем случае не заменяют глубокого знания и понимания предмета исследования. Пути формализации и концептуализации могут быть разными, но главным остаётся приращение нового знания. Для того, чтобы методы применить успешно, нужно хорошо знать предмет, который предполагается измерить, изучить, понять.

ЛИТЕРАТУРА

1. Володин А.Ю. Исторические исследования в контексте датаизма: методологический аспект // Вестник Пермского университета. Серия «История». 2023. Т. 4, № 63. С. 135–147.
2. Майер-Шенбергер В., Кукьер К. Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим. М.: Б. и., 2014.
3. Маккарти У. Специальные эффекты: инструменты есть, а где результаты? // Электронный научно-образовательный журнал «История». 2016. Т. 7, вып. 7 (51).
4. Маклюэн М. Галактика Гутенберга. Становление человека печатающего. М.: Б. и., 2005. 496 с.
5. Робертсон С. Различия между цифровыми гуманитарными науками и цифровой историей // Электронный научно-образовательный журнал «История». 2016. Т. 7, вып. 7 (51).
6. Харари Ю.Н. Homo Deus: краткая история будущего. М.: Синдбад, 2018. 496 с.
7. Цифровые гуманитарные исследования: монография / А.Б. Антопольский, А.А. Бонч-Осмоловская, Л.И. Бородкин [и др.]. Красноярск: Сиб. федер. ун-т, 2023. 272 с.
8. Цифровые гуманитарные науки: хрестоматия: пер. с англ. / ред. М. Террас [и др.]. - Красноярск: СФУ, 2017. 351 с.
9. Data-Driven Innovation in the Creative Industries / Ed. by Melissa Terras, Vikki Jones, Nicola Osborne, Chris Speed. Routledge, 2024. 300 p.
10. Debates in the Digital Humanities / M.K. Gold, L.F. Klein (eds.). 2023. University of Minnesota Press, 2023. 520 p.
11. Drucker J. Humanities Approaches to Graphical Display // DHQ: Digital Humanities Quarterly. 2011. Volume 5. Number 1.
12. Grimmer J., Roberts M.E., Stewart B.M. Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press, 2022. 360 p.
13. The Bloomsbury Handbook to the Digital Humanities (Bloomsbury Handbooks). Bloomsbury Academic, 2022. 512 p.

Журнал «Современная наука: актуальные проблемы теории и практики»