DOI: 10.19181/4m.2021.53.4
EDN: LUWLVB

А. В. Сапонова, С. П. Куликов
(*Москва*)

ИНТЕГРАЦИЯ ОПРОСНЫХ ДАННЫХ И ЦИФРОВЫХ СЛЕДОВ: ОБЗОР ОСНОВНЫХ МЕТОДОЛОГИЧЕСКИХ ПОДХОДОВ[1]

Цель настоящей статьи – рассмотреть основные методологические подходы к интеграции опросных данных и цифровых следов, которые применяются в социологических исследованиях. В работе обсуждается методологическая дискуссия о месте больших цифровых данных в концептуальном аппарате социальных наук. Предпринимается попытка проблематизировать практику интеграции данных опросов и цифровых следов через концепцию «реактивного – нереактивного» измерения. Обозначаются возможные функции цифровых следов (на примере данных социальных медиа) при встраивании в дизайн исследования. На основе трех ведущих исследовательских направлений (изучения медиапотребления, медиаэффектов и электорального поведения) были продемонстрированы общие методологические принципы интеграции данных разной природы, также обозначены возможные перспективы развития

Анастасия Владимировна Сапонова – преподаватель, аспирантка кафедры анализа социальных институтов, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. E-mail: asaponova@hse.ru
Сергей Павлович Куликов – аспирант кафедры анализа социальных институтов, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия. E-mail: spkulikov@hse.ru

[1] Авторы выражают глубокую благодарность Инне Феликсовне Девятко за ценные комментарии к тексту работы.

этих подходов. В статье обсуждается широкий круг методологических вопросов: проблемы валидности связывания данных, потенциальные угрозы валидности цифровых следов, возможности по совершенствованию опросного инструментария, обогащению данных, поиску новых валидных индикаторов социально-политических процессов и кросс-валидации результатов исследований. Отдельно рассматриваются практики интеграции административных данных.

Ключевые слова: интеграция данных; связывание данных; большие данные; нереактивные методы; цифровые следы; опросные данные

Введение

Цифровизация способствовала значительному увеличению типов и количества данных, доступных для исследователей. С одной стороны, эта диверсификация открыла новые возможности для обогащения и кросс-валидации данных, с другой – актуализировала проблему интеграции данных, состоящую из целого спектра сложных методологических задач. Прежде всего к таким задачам относится интеграция (связывание) разных типов данных, полученных из разных источников. Среди всего многообразия социологических типов данных основной корпус работ в области интеграции данных сосредоточен на связывании двух типов – опросных данных и цифровых следов (больших цифровых данных), поэтому в настоящей статье предлагается систематический обзор подходов, которые используются для интеграции этих типов данных.

В настоящей статье мы сначала кратко рассмотрим методологическую дискуссию о роли больших цифровых данных (цифровых следов) и их соотношении с традиционными подходами к социологическим исследованиям. Далее, обозначив подход к различению природы данных (реактивных и нереактивных), мы определим общие принципы интеграции данных разных типов, а также постараемся критически оценить возможности

и ограничения процесса сбора, обработки и анализа цифровых данных. Основные методологические проблемы, возникающие при интеграции, а также формирующиеся подходы к их решению будут рассмотрены в статье на примере нескольких тематически сформировавшихся исследовательских направлений: измерения медиапотребления, медиаэффектов и электорального поведения. Также кратко будет охарактеризован потенциал использования административных данных в исследовательской практике.

Большие цифровые данные как новый источник данных

С появлением Интернета в результате развития информационных технологий компьютеры, мобильные телефоны, роботы, голосовые помощники, умные дома и даже некоторая бытовая техника оказались связаны в единую глобальную сеть обмена информацией. Стремительный успех, открытость и расширяющаяся доступность технологии привлекла внимание ученых и исследователей по мере подключения к Интернету научных вычислительных центров и университетов. В процессе стремительного развития коммуникационных технологий, появления стандарта web 2.0, с изобретением социальных медиа исследователи социальных наук устремились изучать новое исследовательское поле, создаваемое усилиями специалистов в области информационных технологий и энтузиастов. В Интернете увидели возможность найти источник для обнаружения общественных законов и закономерностей в социальных науках, «социальный телескоп», позволяющий наблюдать за поведением большого количества людей [1], и «виртуальную лабораторию», ранее доступную только представителям естественных наук, открывающую доступ для широкого применения эксперимента [2].

В процессе увеличения возможностей использовать в научно-исследовательских целях онлайн-данные часть исследователей

стали отрицать не только необходимость использовать традиционные методы, прежде всего опросы, но и необходимость выдвигать теоретически фундированные гипотезы и, следовательно, применять теорию как таковую. В 2008 г. о «конце теории» заявил К. Андерсон, опубликовав статью *The End of Theory* в журнале *Wired* [3]. Ключевая мысль текста сводится к следующему: с появлением больших цифровых данных необходимо внести изменения в проведение научного исследования, которое может быть выражено в виде циклического последовательного процесса по: 1) выдвижению гипотез, 2) теоретическому моделированию процесса и 3) непосредственной проверке гипотез экспериментом, путем исключения пункта о применении теории в процессе исследования.

Согласно К. Андерсону, если до появления больших данных исследователям было необходимо прибегать к выдвижению гипотез и построению дедуктивно-номологических моделей объяснения из-за нехватки данных и ограничений вычислительных возможностей, то в настоящий момент достигнута возможность обрабатывать большие объемы информации оперативно путем сугубо индуктивного поиска закономерностей без необходимости опираться на теорию и модели объяснения, что кратко может быть выражено словами «корреляции достаточно» [3]. В силу того, что у компаний и государственных организаций накоплены большие данные, не нужно искать причинно-следственные связи и объяснять в принципе наличие связи с точки зрения теории, достаточно лишь обнаружить между явлениями или процессами корреляционную связь.

Тезис К. Андерсона в пользу непосредственного анализа больших данных без необходимости применения концептуальных моделей породил важную дискуссию в научном сообществе [4]. В рамках дискуссии д. бойд и К. Крауфорд критически высказались о способности больших данных принести вклад в научное исследование при учете позиции К. Андерсона, утверждая следующее:

– сами по себе большие данные не являются знанием, без контекста цифры и записи об изучаемом эмпирическом объекте мало о чем говорят, однако способ сбора и обработки больших данных может внести изменения в научное исследование;

– большие данные не отвечают требованиям репрезентативности, некоторые статистические методы могут обнаружить в больших данных взаимосвязи, которые трудно или практически невозможно объяснить;

– большие данные из-за регистрации всего и вся сложно структурированы и из-за этого мало пригодны для научного анализа, в отличие от «малых данных» – данных, структурируемых самими исследователями для целей исследования;

– сами по себе цифры в больших данных не могут говорить за себя – помимо математического описания больших данных важно иметь представление и о способе получения данных;

– большими данными распоряжается небольшой круг лиц, сами данные собираются неэтично [5].

Отмечая наличие у исследователей надежды на прорыв в области анализа больших данных в социальных науках, К. Губа выделяет два подхода к изучению больших цифровых данных в социологии [6]. Первый подход заключается в продолжении развития эмпирико-ориентированной социологии в сторону доказательной социальной науки (forensic social science). С одной стороны, социология стремится удовлетворить запрос на поиск наиболее важных закономерностей, с другой стороны – стремится эти закономерности осмыслить и обосновать [7]. Вторым подходом является стремление перенаправить усилие социологов не на поиск закономерностей и каузальных объяснений, а на описание данных [8]. Стремление отказаться от фокуса на выявление причинно-следственных связей объясняется неравенством возможностей по доступу к данным у социологов по сравнению с аналитиками и маркетологами в условиях «знающего капитализма» [9], «кризисом измерения» [10] и необходимостью развития социологической теории в

цифровую эпоху [11]. Оба подхода акцентируют важность использования больших цифровых данных в исследованиях. Интегрируя их с уже накопленными массивами «малых данных» [12] в социальных науках (в частности, опросных данных), исследователи могут раскрыть потенциал больших цифровых данных. Интеграция данных представляет собой обширный комплекс методологических задач, среди которых: связывание данных разными способами, изучение мотивации респондентов делиться персональной информацией, кросс-валидация параметров, измеренных на разных типах данных, разработка этических норм, связанных с исследованием цифровых следов.

Для обогащения и интеграции с данными опросов чаще всего используют три типа данных – параданные, административные данные и цифровые следы. Под последними, как правило, понимают обширный спектр условно нереактивных (малореактивных) данных о зарегистрированном онлайн-поведении (онлайн-транзакциях, медиапотреблении, геолокации, активности в социальных медиа и т.д). В рамках обзора мы рассмотрим практику интеграции опросных данных и цифровых следов, а также вкратце опишем возможности использования административных данных. Интеграцию данных опросов и параданных, которая представляет собой отдельное и крупное методологическое направление, развивающееся ни одно десятилетие, мы оставим за рамками этой работы, так как оно требует отдельного анализа.

Работы, которые сосредоточены непосредственно на теоретических и методологических особенностях связывания данных, в литературе иногда обозначают как *linkage studies* [13]. Несмотря на относительную новизну этого направления, в литературе уже предпринимаются первые попытки систематизировать возможные методологические подходы в рамках интеграции реактивных и нереактивных данных [14]. Помимо ряда традиционных методологических вопросов (например, насколько и по каким признакам смещена группа, которая готова добровольно делиться

«обогащенной» персональной информацией с исследователями), которые возникают в связи с интеграций данных, исследователи отмечают также правовые и этические. Так, отдельно изучаются процесс получения согласия респондентов на использование их нереактивных данных, их мотивы согласия и/или несогласия делиться персональными данными [15]. Как правило, обозначенные выше вопросы касаются объединения опросных данных и цифровых следов.

В настоящем обзоре не ставится цель предложить новую типологию методологических подходов к интеграции данных, а предпринимается попытка обозначить основные используемые подходы в исследовательской практике, их возможности и ограничения, а также вписать это новое методологическое направление в концепцию «реактивного – нереактивного» измерения, которая подробно будет рассмотрена ниже.

Реактивные и нереактивные данные: определение

Представления о реактивности и нереактивности впервые обстоятельно были изложены Юджином Уэббом и его коллегами в книге «Незаметные меры: нереактивное исследование в социальных науках». Вводя в методологический оборот новую эпистемическую дилемму, они формулируют термин «нереактивные меры» (unobtrusive measures)[1]. Нереактивное измерение определяется как измерение, которое не требует от испытуемого ни участия в исследовании, ни, что, возможно, более важно, осознания факта участия в исследовании [16]. Такое измерение противопоставляется «реактивному» измерению – с использованием классического

[1] На русский язык “unobtrusive” дословно переводится как «ненавязчивый», «незаметный» или «малозаметный», но в научной литературе устоялся перевод этого термина как синонима слова “nonreactive” – «нереактивный».

методологического арсенала социальных и психологических наук (интервью, опросы, эксперименты, фокус-группы и др.), которые основывались на непосредственном и осознанном вовлечении исследуемого в исследовательские процедуры.

Таким образом, в качестве «нереактивных» можно обозначить целый класс данных, полученных в результате такого измерения[1]. Уэбб и его коллеги исходно выделили три типа нереактивных данных – наблюдение (observation), физические следы (physical traces) и архивы (archives) [16], в последнюю группу входят также документы, дневниковые записи, фото- и видеозаписи.

В 1960-х гг. как в социальных, так и в психологических науках доминировали реактивные методы измерения, а нереактивные методы (малозаметное наблюдение, сбор и анализ документов) рассматривались как возможный способ избежать «эффекта реактивного измерения» [16; 17; 18]. Важно отметить, что выделение класса нереактивных методов изначально рассматривалось не как равноправная альтернатива методам реактивным, а как необходимое к ним дополнение [16; 19]. Однако, как отмечают исследователи, несмотря на потенциал, описанный авторами, а также интерес профессионального сообщества к теме, широкого эмпирического применения эта концепция не получила [20; 21].

Среди причин такого «парадигмального неуспеха» называют слабую концептуализированность нереактивных индикаторов (что служит ошибкой при их измерении)[2] и методическую слабость предложенной Уэббом и его коллегами классификации, которая носила описательный характер (упорядочивала разные типы нереактивных данных), но не предоставляла критериев или ориен-

[1] В настоящем обзоре под терминами «метод» и «измерение» понимаются способы сбора данных. Так, дихотомия «реактивные – нереактивные данные» основывается на различении методов, с помощью которых эти данные были добыты.

[2] Нельзя также сказать, что нереактивное измерение в достаточной мере концептуализировано в настоящее время.

тиров для выбора при разработке собственного инструментария [19; 20]. Важно отметить также и влияние такого фактора, как отсутствие четко сформулированных этических норм и правил, регламентирующих использование нереактивных методов измерения. Единственной рекомендацией до 1980-х гг. было представление данных нереактивного измерения в агрегированном виде [20]. Однако, несмотря на указанные недостатки, дихотомия «реактивный – нереактивный» по-прежнему признается удачной для классификации методов социологического измерения и рассматривается как важный признак, отражающий специфику онлайн-исследований [20].

Ренессанс методологических дискуссий по нереактивному измерению произошел после начавшегося процесса цифровизации и глобальной фиксации онлайн-поведения. Одним из первых на возрастающий объем онлайн-информации обратил внимание Р. Ли, который описал цифровые данные как новый тип нереактивных данных [22]. Он же предложил другую классификацию данных, основанную не на разной физической природе, а на роли исследователя в сборе и обработке нереактивных данных, а именно – различать данные на: «найденные» (found), «собранные/зарегистрированные» (captured) и «извлеченные» (retrieved) (см.: [20]). По сравнению с дихотомией «реактивный – нереактивный» в основе этой классификации артикулируется характер активности исследователя (и возможное влияние методов сбора данных на ошибку измерения).

Некоторые исследователи указывают на перманентный рост числа цифровых следов, а также их фрагментированность и неоднородность (см., напр.: [23]). В современной литературе основное внимание уделяется цифровым следам как основному типу «нереактивных» или «малореактивных» данных, однако ранее вполне активно рассматривались и другие виды офлайн-следов – пыль на книжных полках как индикатор объема чтения, число выброшенных бутылок как способ измерить уровень потребления

алкоголя и т.д. [17] – именно физические следы и носители легли в основу концепции «нереактивности».

Граница между реактивными и нереактивными методами не столько строга. Так, существует возможность обогащать реактивные исследования использованием нереактивных или, как минимум, не зависящих от сознательного контроля субъектов, методов измерения (например, айтрекинг, измерение пульса и т.д.) [24]. И наоборот, нереактивное измерение в социальных медиа перестает быть таковым благодаря влиянию таких факторов, как, например, стремление к социальному одобрению, использование стратегий саморепрезентации, которые вносят определенную «реактивность». Более того, стоит учитывать и тот факт, что аудитория может осознавать, что потенциально является объектом внешнего интереса (и не только исследовательского). Таким образом, принимая во внимание эти факторы (как возможные угрозы валидности), применительно к онлайн-среде корректнее будет говорить о «малой реактивности» измерения, а цифровые данные, полученные в результате такого измерения, предлагается называть «малореактивными» [25].

В целом проблема «реактивности» чаще артикулируется психологами, вероятно – в силу более высокой академической культуры[1]. Также можно отметить и низкую степень изученность концепции «нереактивности» российскими социологами. Г. Николаенко и А. Федорова отмечают, что к «нереактивности» как к исследовательской стратегии российские исследователи не обратились ни в период первых волн интереса (в 1970-е и 1990-е гг.), ни в настоящее время, когда очевидно растет внимание к изучению цифровых данных [26].

[1] Регулярно публикуются статьи о необходимости применения нереактивных методов в качестве дополнительного средства оценки психологических конструктов (см., напр.: [18]).

Несмотря на существование алармистского дискурса о «меняющейся научной парадигме» и «новом повороте» в социальных науках [27], в целом исследователи сходятся во мнении, что разработки в области больших данных вряд ли заменят сбор данных реактивными методами, но, вероятно, их дополнят и расширят спектр применяемых методов исследования [28]. «Нереактивность» как концептуальное различение важно артикулировать при обсуждении методологических вопросов интеграции, где одним из важных вопросов является различие между типами связываемых данных, а также способами их получения.

Общие принципы интеграции данных в социальных науках

Современный научный дискурс о «нереактивном» измерении преимущественно связан с областью вычислительной социальной науки (computational social science[1]). Параллельно развивается использование нереактивных данных в области цифровой этнографии [26; 29; 30; 31].

Исторически сложившееся доминирование в социальных науках опросных методов неоднократно критиковалось [16; 17]. Среди основных недостатков доминирующего методологического подхода называются низкая надежность данных, основанных на самоотчете о поведении, установках и т.д. (self-reports), растущая доля неответов [32; 33; 34], а также влияющие на ошибку измерения факторы социальной желательности. Среди основных пре-

[1] Этот термин в 2009 г. ввели в широкий оборот Д. Лезер (Северо-Восточный университет, США) и его коллеги из ряда других американских университетов. В своей статье для журнала Science они характеризуют «вычислительную социальную науку» как научную область, которая обладает возможностями сбора и анализа данных в масштабе, который может выявить закономерности индивидуального и группового поведения [35].

имуществ нереактивных методов принято называть возможности изучения труднодоступных групп и сенситивных тем, а также некоторые технические и инструментальные преимущества – возможность ретроспективного анализа, сокращение финансовых издержек [24]. Не все суждения поддаются вербализации, а припоминание поведения и обсуждение сенситивных тем проблематичны, а также обусловлены временем и местом проведения интервью [19].

С другой стороны, ограничения возникают на этапе оценки обоснованности применения нереактивных методов, в частности при оценке внутренней и внешней валидности измерения. Например, каким образом результаты анализа социальных медиа (как одного из нереактивных методов) могут быть экстраполированы на более широкую генеральную совокупность, как соотносятся измеряемые конструкты и латентные переменные, каким способом организуются неструктурированные цифровые данные [18] – эти вопросы пока не получили внятных ответов в методологии социальных наук.

Однако не только корректное измерение латентных конструктов представляет трудности, цифровые следы прежде всего демонстрируют низкую валидность в измерении базовых социально-демографических параметров. Так, в докладе *Pew Research Center* отмечается, что в результатах *Google Consumer Survey* (GCS) предполагаемый пол (вычисленный предположительно на основе алгоритмов) соответствует заявленному респондентом в рамках опроса примерно в 75% случаев, а возрастные категории совпадают примерно в 44% случаев [36], что существенно снижает валидность и надежность данных, а также ограничивает возможности многомерного анализа [28].

Предпринимаются попытки решить эту проблему с помощью исследований, в рамках которых реконструируется социально-демографический портрет, в частности – проводятся эксперименты по определению географического местоположения (локализации)

пользователя на основе пользовательской информации и параданных [37; 38; 39]1.

Активно обсуждавшая с 1950-х гг. идея совмещения разных исследовательских методов и техник с целью проверки надежности, валидности или же их взаимной интеграции и обогащения (см., напр.: [42; 43; 44]) приобретает еще большую актуальность в дискуссии о возможностях цифровых методов исследования в социологии.

В качестве возможного методологического решения предлагается связывание данных (data linking). С одной стороны, подход может использоваться в целях валидизации и повышения уровня надежности измерения, с другой – он позволяет оптимизировать инструментарий исследования, например – методику опросов (снижать нагрузку на респондента, заполняя некоторые ответы исходя из нереактивных данных). Это связывание может происходить на *агрегированном* или *индивидуальном* уровнях.

Интеграция на агрегированном уровне предполагает сравнение индикаторов, измеренных на разных типах данных. Интеграция этого типа позволяет оценить параметры и сравнить их на макроуровне, не позволяя спускаться до анализа микрогрупп или конкретного индивида. Иными словами, данные не обязательно должны принадлежать одному респонденту. Индикатором для связывания в таком случае может выступать временной период или географическое положение [14]. Например, можно сравнить оценку деятельности политика за определенный период, полученную в рамках опросов, и индикатор семантической окраски сообщений

[1] Twitter* стал главным источником сообщений для разработки алгоритмов по определению геолокации пользователей, разметки их места жительства, работы. Подобные алгоритмы строятся преимущественно на основе анализа метаданных пользователей, извлечении географических названий из сообщений или идентификации геолокации через установление связей с другими пользователями [40]. Однако общим недостатком подобных проектов можно назвать низкую воспроизводимость результатов и непрозрачность используемых алгоритмов [41].

* Социальная сеть заблокирована в РФ 04.03.2022.

из социальных медиа с упоминаем этого политика за аналогичный период [45] или же сопоставить онлайн-активность жителей района с уровнем их декларируемой протестной активности [46]. Основная сложность в данном типе связывания заключается в построении эквивалентных показателей на разных типах данных, сравнение которых было бы корректным [14].

Связывание на индивидуальном уровне, напротив, предполагает связывание разных типов данных, принадлежащих одному респонденту. Например, связывание ответов респондента и его активности в социальных медиа. Такое связывание, в свою очередь, может происходить *детерминированно* (deterministic matching method) – на основе единого индикатора для каждого респондента, который содержится в используемых массивах. Или же *вероятностно* (probabilistic matching method), когда индикатор для связывания устанавливается путем вычисления вероятности на основе параметров, которые содержатся в связываемых массивах [47]. Такой вид связывания позволяет объединять неструктурированные данные. В целом связывание на индивидуальном уровне теоретически представляет возможность выходить на более детальный уровень анализа, однако оно представляется более сложным как с технологической, так и с этической точек зрения.

С. Штир и его соавторы предлагают двумерную классификацию методологических подходов к интеграции данных, где связывание может являться либо частью дизайна исследования с самого начала и данные для интеграции собираются исходя из установленных задач (“ex ante” – «до»), либо связывание происходит постфактум (и сбор происходит независимо друг от друга) на основе собранных данных (“ex post” – «после»). С другой стороны, связывание этими двумя способами может происходить на уровне агрегированном, индивидуальном, а также – на уровне корпоративных (или организационных) акторов (правительств, политиков, и других организаций) [14]. Примеров связывания на последнем уровне можно привести немного [48; 49], однако

это направление можно назвать перспективным с учетом повышения медиаактивности различных государственных структур (и их представителей), а также коммерческих и некоммерческих организаций.

М. Шоббер и его соавторы выделяют четыре основных вопроса, связанных с интеграцией разных типов данных:

а) как участники (респонденты, пользователи социальных сетей) определяют свою деятельность, какие интенции в нее вкладывают;

б) каков характер и природа данных, полученных в ходе опроса, и сообщений в социальных медиа;

в) насколько выводы, полученные на таких данных, могут быть валидными;

г) разработка практических и этических стандартов использования этих данных [50].

Применение нереактивных данных в социальных исследованиях проблематизирует не только методологическую сторону этого вопроса, но требует решения ряда этических и юридических вопросов [51]. Таких как сбор, хранение и последующая обработка данных пользователей, необходимость получения информированного согласия на участие в исследовании, а также этичность создания фиктивного образа исследователя для интеграции в закрытые сообщества [26]. Также остается нерешенной и проблема сохранения нереактивных данных для анализа – как и физические носители, цифровые следы могут быть частично «повреждены» (т.е. удалены) [20], что может влиять как на качество данных, так и на логику получаемых выводов. Как указывалось выше, нереактивность в цифровой среде является таковой лишь условно. При этом возможные угрозы валидности нереактивных данных до сих пор систематически не исследованы, в то время как изучением угроз валидности опросных данных основательно занимались с 1960-х гг. [52].

Проблематика использования данных социальных медиа

Данные из социальных медиа, которые занимают значительное место в классе цифровых данных, могут выполнять в связывании данных двойную функцию. С одной стороны, социальные медиа располагают большими массивами малореактивных данных, которые могут извлекаться с помощью веб-скрейпинга и связываться с респондентом (если он при проведении опроса предоставил информированное согласие на использование этих данных) на индивидуальном уровне, а также возможны ретроспективный сбор и анализ пользовательских данных на агрегированном уровне. С другой стороны, социальные медиа могут выступать в качестве «точек входа» для привлечения участников с целью проведения дополнительных исследовательских процедур, такой площадкой по сбору информации для их последующего привлечения часто выступает Facebook* [53]. В качестве одного из самых широко известных таких онлайн-экспериментов можно отметить исследование Михаила Косиньского и его соавторов с использованием приложения *myPersonality*[1] на Facebook*, которое давало участникам обратную связь по их личностным качествам [54]. С помощью классического теста «Большая пятерка» исследователи собрали данные о таких личностных качествах респондентов, как интеллект, удовлетворенность жизнью, употребление психоактивных веществ и др. Приложение также собирало данные с профилей пользователей (возраст, пол, количество друзей, лайки на страницах). Среди ограничений эксперимента отмечают, например, возникающие сильные смещения вследствие самоотбора – даже при большом количестве участников [55].

*Facebook принадлежит компании Meta, которую признали экстремистской, запрещен в РФ 02.03.2022.

[1] Приложение было разработано исследователями. Всего «тест» прошли 400 0000 пользователей, 40% из них дали согласие на сбор метаданных с их страниц.

Использование данных из другого аналогичного по дизайну эксперимента (с использованием приложения *This Is Your Digital Life* на Facebook*) обернулось громким скандалом в 2018 г. – на протяжении нескольких лет британская консалтинговая компания *Cambridge Analytica*, получавшая данные от создателя приложения, использовала их для настройки таргетированной политической рекламы [56]. Компания также покупала отдельные массивы данных (с информацией об онлайн-транзакциях, истории поиска и т.д.), которые уже после («ex post») связывались с профилями пользователей на Facebook*. На примере этого кейса по масштабной интеграции (и коммерциализации) собранных в социальных сетях поведенческих цифровых следов и опросных данных хорошо виден спектр этических и юридических проблем, которые предстоит решить не только на уровне исследовательских конвенций, но и законодательного регулирования.

Помимо обсуждавшихся ранее технологических проблем, решения требует и вопрос о поиске общей теоретической и концептуальной рамки для интеграции разных типов данных. Решение в этой области ищется в направлении определения природы данных и поиске общих оснований для интеграции. Очевидно, что методология извлечения и использования данных из социальных медиа имеет ряд принципиальных отличий, однако в поиске общего знаменателя для интеграции некоторые исследователи предлагают рассматривать данные из социальных медиа как некоторый аналог опросных данных. Так, например, предлагается рассматривать данные из социальных медиа как ответы опроса по неслучайной (и изменяющейся в динамике) выборке из добровольных участников, которые выборочно отвечают на вопросы [57]. Однако, на наш взгляд, такие попытки – за счет некой искусственной унификации данных решить проблему валидности связывания – не способствуют формированию методологических программ и стандартов по интеграции данных.

Некоторые исследователи предлагают рассматривать поведение пользователей социальных медиа не на агрегированном уровне (ибо не совсем понятно, составляют ли они из себя цельную социальную группу), а на индивидуальном – где можно наблюдать, например, повторяющиеся действия, не делая при этом ложные обобщения [57]. Довольно часто связывание данных и их интеграция подменяются поиском корреляций в разных типах данных, не имеющих под собой теоретического обоснования или должной концептуализации. Вместе с этим представляется необходимым учитывать и ограничения статистических методов анализа, используемых для разных типов данных, – здесь существует, как пример, риск ложных корреляций [38]. В проектах по интеграции опросных данных и цифровых следов большое значение играют выбранные методы фильтрации контента, способы классификации текста (в частности – коэффициенты сглаживания), которые могут влиять на содержательные результаты [58].

Вопрос о валидности измерения настроений в социальных медиа по образу и подобию опросов общественного мнения сравнительно активно обсуждался в 2013–2014 гг. на волне некорректного предсказания сервисом *Google Flu Trends* вспышки гриппа (алгоритм преувеличил размер эпидемий на 50%) [59]. Основная идея проекта заключалась в том, что по участившимся поисковым запросам с упоминанием слова «грипп» возможно судить о потенциальной вспышке заболевания[1]. Алгоритм был настроен на отслеживание степени распространенности гриппа и прогнозирование роста заболеваемости в реальном времени – на две недели раньше, чем это могли сделать специализированные медицинские учреждения, которые делают прогнозы исходя только из зарегистрированных случаев болезни.

[1] Цифровой след здесь понимается как поведенческий «отпечаток» некоторого физиологического процесса – когда люди болеют гриппом, многие ищут информацию, связанную с гриппом в Google, пытаясь узнать больше о болезни, найти схемы для самолечения и т.д.

Кейс породил среди исследователей обсуждения о надежности и прозрачности предсказательных алгоритмов, основанных на больших данных, а также систематическом воспроизведении результатов. Именно воспроизводимость (replicability) является основной проблемой в прогностических моделях, которые строятся на основе нереактивных данных.

В отдельных случаях предпринимались попытки использовать анализ нереактивных данных в качестве, предположительно, более валидного метода, то есть с целью более точного измерения одного и того же латентного конструкта. Например, степень расовой сегрегации в группе студентов оказалась выше по результатам анализа их френдленты, тогда как основанные на самоотчетах данные о социальных взаимодействиях студентов говорили о более гетерогенной коммуникации [60].

Таким образом, несмотря на существующий обширный функционал по сбору и обработке «малореактивных» цифровых данных, основным ограничением остается недостаток в определении природы данных социальных медиа, их возможности быть встроенными в построенные на принципах интеграции исследовательские программы и – в последующем – в модели социологического объяснения.

В целом тематика исследований в области связывания данных довольно разнообразна, однако на настоящий момент сложилось несколько основных тематических направлений: измерение медиапотребления, медиаэффектов и электорального поведения. Охарактеризуем далее кратко каждое из них.

Основные тематические направления интеграции данных

Измерение медиапотребления

В качестве одной из давних исследовательских традиций, где предпринимались попытки интеграции данных реактивных

и нереактивных (в основном с целью кросс-валидации), отметим изучение медиапотребления. Общий недостаток классических медиаметрических техник, которые применялись с 1930–1940-х гг. в США, в том числе в панельных исследованиях – дневниковых записей, техники *day after recall* (телефонного интервью на следующий день после эфира), – заключался в угрозе валидности и надежности при припоминании и ретроспективной оценке действий. Так, результаты предыдущих исследований говорят о низкой степени точности такой самофиксации [61; 62; 63; 64; 65; 66; 67]. Однако в силу отсутствия альтернативных способов более «объективного» измерения самоотчеты долгое время оставались единственным способом измерения медиапотребления[1].

С экспоненциальным ростом числа каналов коммуникации и увеличением объема медиапотребления применение исследовательских методик, основанных на субъективном самоотчете, вызывает много вопросов относительно корректности такого измерения [14]. Один из последних метаанализов в этой области, основываясь на 106 работах с рассчитанным эффектом, показывает низкий уровень корреляции самоотчетов и нереактивных данных о зарегистрированном медиапотреблении [69], и, как следствие, авторы делают вывод о низкой надежности самоотчетов. Обзоры в этой области свидетельствуют о том, что респонденты преувеличивают интенсивность использования Интернета, не могут точно вспомнить факты посещений конкретных веб-сайтов и их частоту, а также склонны завышать количество посещений новостных онлайн-сайтов и потребления контента [14; 70]. Систематические смещения выражаются в завышении объема потребления

[1] За исключением измерения потребления телевидения, которое с конца 1940-х гг. измерялось с помощью пиплметров параллельно с дневниковыми исследованиями, а с конца 1980-х гг. – уже преимущественно пиплметрами [68]. Очевидные недостатки и смещения в дневниковых записях долго уравновешивали издержки пиплметрического измерения, которое все еще фиксирует только факт включенного телеканала, но не его фактическое потребление.

Интернета [71], телевидения [72; 73] или частоты пользования мобильным телефоном [74]. Эксперименты по валидации оценок медиапотребления ставились еще до широкого распространения Интернета [75; 76; 77].

Измерения медиавоздействия

Другое довольно узкое исследовательское направление продолжает долгую традицию эмпирико-функционалистской школы, связанную с интеграцией данных опросов и некоторых условно объективных индикаторов (данных медиапотребления) с целью измерения эффекта медиавоздействия [78; 79; 80]. Фактически предпринимаются попытки фиксации на микроуровне воздействия на респондентов экзогенных факторов в виде источников информации. Классическим примером (скорее даже прототипом) в этой области является сделанное на агрегированном уровне исследование М. Мак-Комбса и Д. Шоу, которое легло в основу теории формирования повестки дня (agenda-setting theory), где авторы сопоставляли данные опроса о проблемах, которые люди номинируют как важные, и темы, которые освещаются в СМИ [81].

Чаще всего подобные проекты делаются в русле электоральной социологии и исследований политических предпочтений и ориентаций в целом. Здесь возможны различные вариации исследовательских дизайнов, однако чаще всего в опрос включается расширенный блок вопросов по медиапотреблению, в дальнейшем результаты опросов связываются с результатами контент-анализа. Для каждого респондента рассчитывается «показатель воздействия» (measure of exposure), где взвешиваются данные из самоотчета о медиапотреблении и заметности какой-то темы в СМИ (например, заметность политической партии, поднимаемых в СМИ проблем, оценок политических действий). Результаты исследований в этой области говорят о существовании корреляции между общественным мнением и проводимыми в СМИ политическими кампаниями [79; 82; 83; 84]. Использование таких методик возможно, если

за основу берутся линейные представления о медиавоздействии, где сообщение, предположительно, воздействует напрямую на реципиентов. Однако для измерений, где медиавоздействие концептуализируется несколько сложнее, например – через двухступенчатый поток информации (теория П. Лазерсфельда), данные опросов необходимо связывать не только с данными о медиапотреблении, но и с информацией о социальных контактах (а также политических предпочтениях респондентов), что требует разработки и проведения чуть более сложных по конструкции методологических экспериментов.

Исследования электорального поведения

Другой крупный корпус литературы по интеграции данных сосредоточен в области электоральных исследований. Примером таких исследований является сравнение результатов опросов о поддержке тех или иных политических сил с размерами онлайн-аудитории групп поддержки в период избирательной кампании [85; 86], географией пользователей социальных сетей, комментирующих политические процессы [87], тематикой и тональностью постов, распространяемых в социальных медиа [45; 88]. Благодаря интеграции опросов и результатов выборов было обнаружено, что существует корреляционная связь между количеством лайков под постами политических лидеров и результатами голосования [89; 90; 91], а позднее была разработана модель, демонстрирующая возможность предсказать распределение голосов, схожее с результатами электоральных опросов, по приросту уникальных пользователей, поставивших отметку «мне нравится» под контентом кандидата на выборах [92]. В одних исследованиях утверждается, что заметность и обсуждаемость политика в «Твиттере» статистически значимо коррелирует с количеством голосов за этого кандидата [93], в других – что существует связь между тональностью сообщений в «Твиттере» и результатами выборов в президенты США [87].

Результаты эксперимента по измерению отношения пользователей «Твиттера» к президенту Бараку Обаме и интеграции с данными аналогичных опросов общественного мнения [45] показали, что на краткосрочных отрезках индикаторы одобрения имеют разные значения, но более схожи в долгосрочной перспективе. Неоднократно предпринимались попытки предсказать поведение избирателей и их партийные предпочтения, используя и другие виды цифровых следов и данных опросов. Так, совместив информацию о десктопном и мобильном потреблении контента пользователями, имеющими право голосовать на федеральных выборах в Германии в 2017 г., с данными опроса ($N = 2000$), было показано, что онлайн-активность не предсказывает результаты голосования [94]. Исследования, где поведение пользователя в социальных сетях рассматривается как предиктор политических процессов и/или индикатор политического поведения (см., напр.: [93]), критиковались как редукционистские и невалидные [95] в силу отсутствия стандартизированных подходов к сбору и анализу цифровых данных из социальных сетей и ненадежности такого индикатора, как упоминание политического лидера или политической партии, а также тональности сообщения с подобным упоминанием в социальных медиа для объяснения электорального офлайн-поведения.

Возвращаясь к теме изучения политических предпочтений в рамках электоральной социологии, стоит указать, что исследования показывают, как пользователи с более высоким уровнем заинтересованности в политике зачастую переоценивают свою активность в публикации и распространении политического контента в социальных медиа, на самом деле больше уделяя внимание потреблению контента и обсуждению его с такими же заинтересованными в политике друзьями и знакомыми по социальной сети [63; 96; 97]. При этом в сравнении с менее заинтересованными в политике пользователями социальных сетей более заинтересованные в политике точнее определяют свое участие в конкретных политических акциях и в целом лучше определяют

те действия, которые являются политическими [63]. Эксперименты с совмещением данных социальных медиа и данных айтрекинга позволили установить, что на постах с политическим контентом пользователи социальных медиа в среднем останавливают взгляд на 4 секунды (что на 25% меньше, чем потраченное время на посты с новостями), переоценивают количество постов, изображений, которые они видели в социальных медиа, и чаще ошибаются в определении типа контента (политический, новостной и т.д.) и источнике контента [97], из чего следует необходимость в процессе опросов больше уделять внимания объяснению респонденту того, что исследователь понимает под политическим контентом и политическим действием.

При описании электората и его политического поведения в Интернете зачастую используется теория «пузыря», или «эхо-камеры» [98], согласно которой интернет-аудитория политических новостных сайтов / социальных медиа идеологически поляризована, представляет из себя гомогенные группы потребителей политического контента, в которых не выражается стремление к идеологическому разнообразию в выборе источников информации. Совмещение цифровых следов, собранных посредством кликстрима (clickstream – сбор и анализ агрегированных данных о посещении сайтов), и опросов эмпирически продемонстрировало ограниченность теории «эхо-камеры»: по результатам исследований сопоставления ответов респондентов и списков посещенных ими сайтов утверждается, что потребители политического контента не представляют из себя гомогенную группу и получают информацию из идеологически разнообразных источников (в том числе неучтенных исследователями), о чем было сложно узнать только лишь из опросов пользователей при помощи вопросов с закрытым списком ответов, но стало возможно благодаря связыванию данных [99].

Несмотря на то, что в фокусе настоящей статьи находится интеграция опросных данных и цифровых следов, далее мы кратко

остановимся на практике использования административных данных и их потенциале для исследовательских проектов, использующих опросные данные.

Практика использования административных данных

Административные данные, собираемые и производимые финансовыми структурами, школами, службами занятости, социального обеспечения и здравоохранения, давно известны социологам как возможный инструмент для кросс-валидации. Так, еще в 1949 г. в рамках так называемого «Денверского исследования валидности» данные местной статистики были сопоставлены с ответами респондентов на фактологические вопросы (о наличии читательского билета в библиотеке, водительских прав и т.д.) [52]. Цифровизация существенно изменила масштаб и разнообразие доступных исследователям административных данных, к ним теперь можно отнести данные мониторинга пользования городским транспортом, системами видеонаблюдения и т.д. Эти новые типы могут рассматриваться как подвид больших данных, подробно критерии такого обобщения обсуждаются исследователями [100; 101]. Среди преимуществ использования административных данных в дизайне исследования отмечают возможность получения информации как для больших выборочных совокупностей, так и для труднодоступных групп населения [102]. Использование административных данных имеет особое значение в лонгитюдных исследованиях, измерении долгосрочных эффектов, где необходим сбор данных о группе (когорте) на протяжении длительного времени [100]. В последнее время предпринимались попытки с помощью привлечения административных данных изучить факторы, влияющие на мобильность мигрантов [103], измерить международную миграцию [104], оценить влияние социальных контактов на политическое поведение [105].

В качестве потенциальных недостатков и рисков инкорпорирования административных данных в исследовательскую практику можно обозначить такие характерные для «найденных» (found) данных черты как угроза качеству и полноте этих данных (отсутствие отдельных записей или ошибки в записях), труднодоступность (закрытость) этих данных. Важным и специфическим методологическим вопросом здесь является и обеспечение конфиденциальности при связывании данных (privacy-preserving record linkage) [106]. Так, для связывания административных данных с другими типами данных (опросными преимущественно) сформирован ряд моделей[1], которые отличаются разделением процессов увязки между участвующими в процессе субъектами (исследователями, организациями) [102]. Одна из таких моделей предполагает, что идентифицируемые административные данные доступны только доверенной третьей стороне (которая проводит привязку), в то время как исследователи получают доступ только к неидентифицированным (закодированным) атрибутивным данным [107]. Другая модель предполагает связывание данных на стороне исследователя. Например, процесс связывания данных в общенациональном британском лонгитюдном исследовании *Understanding Society* осуществляется следующим образом: исследователи передают в государственную организацию массив, содержащий временное ID респондента, и часть его персональной информации – имя, пол, дату рождения и адрес проживания. Организация, со своей стороны, находит записи об этих респондентах, удаляет персональную информацию (оставляя только временное ID) и передает исследователям имеющиеся данные на запрашиваемых

[1] В настоящее время устойчивые практики обмена и связывания административных данных существуют в Норвегии, Финляндии, Швеции; в Великобритании и США они находятся на стадии формирования [99].

респондентов. Далее исследователи интегрируют ответы респондентов и уже анонимизированные административные данные[1].

Помимо решения обозначенных методологических вопросов, возможность интеграции административных данных требует сотрудничества между различными административными образованиями, которые владеют этими данными, и исследователями, а также разработки механизмов управления обмена данными.

Заключение

Как было показано в работе, исследовательское направление по интеграции опросных данных и цифровых следов сформировалось в ответ на потребность в проработке на теоретико-методологическом уровне вопроса о применении цифровых следов в исследованиях, а также в решении задач по кросс-валидации результатов исследования, обогащению данных, оптимизации дизайна исследования. В обзоре были продемонстрированы основные тематические направления, в рамках которых практикуется интеграция данных разных типов. Ниже в таблице суммированы важные характеристики этих направлений.

Так, исследования в области медиапотребления демонстрируют ограничения опросного инструментария в вопросе самооценки медиапотребления, затруднения с припоминанием и т.п. В области изучения медиавоздействия исследователи продолжают работу по разработке подходящих инструментов, способных выделить и оценить силу влияния информационных кампаний в СМИ и новых медиа на общественное мнение. В области электоральных исследований ведется поиск в социальных медиа индикаторов, способных выступить аналогом при оценке способности политических лидеров получить победу на выборах, а также уточняется

[1] Важно отметить, что это связывание осуществляется только при наличии информированного согласия от респондента.

Таблица 1

ВЕДУЩИЕ ТЕМАТИЧЕСКИЕ НАПРАВЛЕНИЯ ИНТЕГРАЦИИ ДАННЫХ

Объект исследования	Типы интеграции		Методологические вопросы	
	Уровень	Тип сбора данных	Решенные проблемы	Перспективы
Медиапотребление	Индивидуальный	Связывание “ex ante”	Валидация ретроспективных самоотчетов о медиапотреблении при помощи данных, собранных регистрирующими медиаповедение программами и устройствами [70; 71; 73]	Повышение точности измерения регистрирующих медиапотребление программ и устройств
Медиавоздействие	Индивидуальный Агрегированный	Связывание “ex ante” и “ex post”	Измерение воздействия на индивида источников информации, расчет показателя воздействия на основе опросных данных и результатов контент-анализа текстов СМИ [78; 79; 80]	Использование нелинейных моделей измерения информационного воздействия
Электоральное поведение	Индивидуальный Агрегированный	Связывание “ex ante”	Прогнозирование электорального поведения и партийных предпочтений в социальных медиа на основе индикаторов вовлеченности (лайков, репостов), тональности публикаций [85; 87; 88]	Разработка валидных индикаторов электорального поведения в социальных медиа; стандартизация методов сбора и анализа данных социальных медиа

вопрос о распространении политической информации в Интернете и предпочтениях пользователей в выборе источников информации. Общим для всех описанных направлений является демонстрация эвристических возможностей применения интеграции опросных данных и цифровых следов при связывании данных. При этом нуждаются в дальнейшей проработке такие проблемы, как выработка стандартов по интеграции данных; разработка эквивалентных показателей, позволяющих сравнивать результаты, полученные на основе опросных данных и цифровых следов и общих переменных, способных обеспечить успешную интеграцию данных как на индивидуальном, так и на агрегированных уровнях связывания.

ЛИТЕРАТУРА

1. *Golder S.A., Macy M.W.* Digital Footprints: Opportunities and Challenges for Online Social Research // Annual Review of Sociology. 2014. Vol. 40, No. 1. P. 129–152. DOI: 10.1146/annurev-soc-071913-043145

2. *Salganik M.J., Watts D.J.* Web-Based Experiments for the Study of Collective Social Dynamics in Cultural Markets // Topics in Cognitive Science. 2009. Vol. 1, No. 3. P. 439–468. DOI: 10.1111/j.1756-8765.2009.01030.x

3. *Anderson C.* The End of Theory: The Data Deluge Makes the Scientific Method Obsolete // Wired. 23.06.2008. URL: https://www.wired.com/2008/06/pb-theory/ (дата обращения: 17.04.2022).

4. *Майер-Шенбергер В., Кукьер К.* Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / Пер. с англ. И. Гайдюк. М.: Манн, Иванов и Фербер, 2014. 240 с.

5. *Boyd d., Crawford K.* Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon // Information, communication & society. 2012. Vol. 15, No. 5. P. 662–679. DOI: 10.1080/1369118X.2012.678878

6. *Губа К.С.* Большие данные в социологии: Новые данные, новая социология? // Социологическое обозрение. 2018. № 17 (1). С. 213–236. DOI: 10.17323/1728-192X-2018-1-213-236

7. *McFarland D. A., Lewis K., Goldberg A.* Sociology in the era of big data: The ascent of forensic social science // The American Sociologist. 2016. Vol. 47, No. 1. P. 12–35. DOI: 10.1007/s12108-015-9291-8

8. *Burrows R., Savage M.* After the crisis? Big Data and the methodological challenges of empirical sociology // Big Data & Society. 2014. Vol. 1, No. 1. DOI: 10.1177/0038038507080443

9. *Thrift N.* Knowing Capitalism. London: SAGE Publications Ltd, 2005. 256 p. DOI: 10.4135/9781446211458

10. *Gane N.* Measure, Value and the Current Crises of Sociology // The Sociological Review. 2011. Vol. 59, No. 2. P. 151–173. DOI: 10.1111/j.1467-954X.2012.02054.x

11. *Ignatow G.* Sociological Theory in the Digital Age. 1st ed. New York: Routledge, 2020. 120 p. DOI: 10.4324/9780429292804

12. *Kitchin R., Lauriault T.P.* Small data in the era of big data // GeoJournal. 2015. Vol. 80, No. 4. P. 463–475. DOI: 10.1007/s10708-014-9601-7

13. *de Vreese C.H. et al.* Linking Survey and Media Content Data: Opportunities, Considerations, and Pitfalls / C.H. de Vreese, M. Boukes, A. Schuck, R. Vliegenthart, L. Bos, Y. Lelkes // Communication Methods and Measures. 2017. Vol. 11, No. 4. P. 221–244. DOI: 10.1080/19312458.2017.1380175

14. *Stier S. et al.* Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field / S. Stier, J. Breuer, P. Siegers, K. Thorson // Social Science Computer Review. 2020. Vol. 38, No. 5. P. 503–516. DOI: 10.1177/0894439319843669

15. *Beninger K. et al.* Understanding Society: How people decide whether to give consent to link their administrative and survey data / K. Beninger, A. Digby, G. Dillon, J. MacGregor // Understanding Society Working Paper Series. 2017. No. 13. 65 p.

16. *Webb E.J. et al.* Unobtrusive measures: nonreactive research in the social sciences / E.J. Webb, D.T. Campbell, R.D. Schwartz, L. Sechrest. Chicago: Rand McNally, 1966. 220 p.

17. *Bouchard Jr T.J.* Unobtrusive Measures: An Inventory of Uses // Sociological Methods & Research. 1976. Vol. 4, No. 3. P. 267–300. DOI: 10.1177/004912417600400301

18. *Hill A.D., White M.A., Wallace J.C.* Unobtrusive measurement of psychological constructs in organizational research // Organizational Psychology Review. 2014. Vol. 4, No. 2. P. 148–174. DOI: 10.1177/2041386613505613

19. *Lee R.M.* Unobtrusive Methods // Handbook of Research Methods in Health Social Sciences / Ed. by P. Liamputtong. Wiesbaden: Springer VS, 2019. P. 491–507. DOI: 10.1007/978-981-10-5251-4_85

20. *Девятко И.Ф.* Инструментарий онлайн-исследований: попытка каталогизации // Онлайн-исследования в России 3.0 / Отв. ред. И.Ф. Девятко, А.В. Шашкин, С.Г. Давыдов; науч. ред. И.Ф. Девятко. М.: OMI RUSSIA, 2012. С. 17–30.

21. *Дудина В.И.* Цифровые данные – потенциал развития социологического знания // Социологические исследования. 2016. № 9. С. 21–30.

22. *Lee R.M.* Unobtrusive Measures in Social Research. Philadelphia, PA: Open University Press. 2000. 192 p.

23. *Kalokyri V. et al.* Integration and Exploration of Connected Personal Digital Traces / V. Kalokyri, A. Borgida, A. Marian, D. Vianna // Proceedings of the ExploreDB'17. Chicago, IL: ACM, 2017. DOI: 10.1145/3077331.3077337

24. *Araujo T., Neijens P.* Unobtrusive Measures for Media Research // The International Encyclopedia of Media Psychology. 1st ed. / Ed by J. Bulck. Hoboken, NJ: Wiley Blackwell, 2020. P. 1–7. DOI: 10.1002/9781119011071.iemp0049

25. *Девятко И.Ф.* Новые данные, новая статистика: от кризиса воспроизводимости к новым требованиям к анализу и представлению данных в социальных науках // Социологические исследования. 2018. № 12. С. 30–38.

26. *Федорова А.А., Николаенко Г.А.* Нереактивная стратегия: применимость незаметных методов сбора социологической информации в условиях web 2.0 на примере цифровой этнографии и big data // Социология власти. 2017. Т. 29, № 4. С. 36–54.

27. *Savage M., Burrows R.* The Coming Crisis of Empirical Sociology // Sociology. 2007. Vol. 41, No. 5. P. 885–899. DOI: 10.1177/0038038507080443

28. *Couper M.P.* Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys // Survey Research Methods. European Survey Research Association. 2013. Vol. 7, No. 3. P. 145–156. DOI: 10.18148/SRM/2013.V7I3.5751

29. *Beaulieu A.* Sociable hyperlinks: an ethnographic approach to connectivity // In Virtual Methods: issues in social research on the Internet / Ed by C. Hine. Oxford: Berg, 2005. P. 183–198.

30. *Hine C.* Internet Research and Unobtrusive Methods // Social Research Update. 2011. No. 61. P. 1–4.

31. *Dirksen V., Huizing A., Smit B.* 'Piling on layers of understanding': the use of connective ethnography for the study of (online) work practices // New Media & Society. 2010. Vol. 12, No. 7. P. 1045–1063. DOI: 10.1177/1461444809341437

32. *De Heer W., De Leeuw E.* Trends in household survey nonresponse: A longitudinal and international comparison // Survey nonresponse. 2002. Vol. 41, P. 41–54.

33. Nonresponse in Social Science Surveys: A Research Agenda / Ed. by R. Tourangeau, T.J. Plewes. Washington, DC: The National Academies Press, 2013. 151 p.

34. *Čehovin G., Bosnjak M., Lozar Manfreda K.* Item Nonresponse in Web Versus Other Survey Modes: A Systematic Review and Meta-Analysis // Social Science Computer Review. 2022. DOI: 10.1177/08944393211056229

35. *Lazer D.et al.* Computational social science / D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, M. Van Alstyne // Science. 2009. Vol. 323, No. 5915. P. 721-723. DOI: 10.1126/science.1167742

36. *Keeter S., Christian L.* A comparison of results from surveys by the Pew Research Center and Google Consumer Surveys. Washington, DC: Pew Research Center, 2012. 30 p.

37. *Graham M., Hale S.A., Gaffney D.* Where in the world are you? Geolocation and language identification in Twitter // The Professional Geographer. 2014. Vol. 66, No. 4. P. 568–578. DOI: 10.1080/00330124.2014.907699

38. *Conrad F.G. et al.* Social Media as an Alternative to Surveys of Opinions About the Economy / F.G. Conrad, J. A. Gagnon-Bartsch, R. A. Ferg, M. F. Schober, J. Pasek, E. Hou // Social Science Computer Review. 2021. Vol. 39, No. 4. P. 489–508. DOI: 10.1177/0894439319875692

39. *Schulz A. et al.* A Multi-Indicator Approach for Geolocalization of Tweets / A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, M. Mühlhäuser // ICWSM. 2021. Vol. 7, No. 1. P. 573–582.

40. *Stock K.* Mining location from social media: A systematic review // Computers, Environment and Urban Systems. 2018. Vol. 71. P. 209–240. DOI: 10.1016/j.compenvurbsys.2018.05.007

41. *Chen Q., Poorthuis A.* Identifying home locations in human mobility data: an open-source R package for comparison and reproducibility // International Journal of Geographical Information Science. 2021. Vol. 35, No. 7. P. 1425–1448. DOI: 10.1080/13658816.2021.1887489

42. *Campbell D.T., Fiske D.W.* Convergent and discriminant validation by the multitrait-multimethod matrix // Psychological bulletin. 1959. Vol. 56, No. 2. P. 81–105. DOI: 10.1037/h0046016

43. *Bouchard Jr T.J.* Field research methods: Interviewing, questionnaires, participant observation, systematic observation, unobtrusive measures // Handbook of industrial and organizational psychology. Vol. 1. / Ed. by M.D. Dunnette. Chicago: Rand McNally, 1976. P. 363–413.

44. *Zeller R.A., Carmines E.G.* Measurement in the social sciences: The link between theory and data. Cambridge; New York: Cambridge University Press, 1980. 198 p.

45. *Pasek J. et al.* Who's Tweeting About the President? What Big Survey Data Can Tell Us About Digital Traces? / J. Pasek, C. A. McClain, F. Newport, S. Marken // Social Science Computer Review. 2020. Vol. 38, No. 5. P. 633–650. DOI: 10.1177/0894439318822007

46. *Климова А.М., Куликов С.П., Чмель К.Ш.* Роль социальных медиа в формировании регионального экологического протеста в России // Мониторинг общественного мнения: Экономические и социальные перемены. 2021. № 6. С. 28–52. DOI: 10.14515/monitoring.2021.6.2024

47. *Shlomo N.* Overview of Data Linkage Methods for Policy Design and Evaluation // Data-Driven Policy Impact Evaluation / Ed by N. Crato, P. Paruolo. Cham: Springer International Publishing, 2019. P. 47–65

48. *Quinlan S. et al.* 'Show me the money and the party!' – variation in Facebook and Twitter adoption by politicians / S. Quinlan, T. Gummer, J. Roßmann, C. Wolf // Information, Communication & Society. 2018. Vol. 21, No. 8. P. 1031–1049. DOI: 10.1080/1369118X.2017.1301521

49. *Karlsen R., Enjolras B.* Styles of Social Media Campaigning and Influence in a Hybrid Political Communication System: Linking Candidate Survey Data with Twitter Data // The International Journal of Press/Politics. 2016. Vol. 21, No. 3. P. 338–357. DOI: 10.1177/1940161216645335

50. *Schober M. F. et al.* Social Media Analyses for Social Measurement / M.F. Schober, J. Pasek, L. Guggenheim [et al.] // Public Opinion Quarterly. 2016. Vol. 80, No. 1. P. 180–211. DOI: 10.1093/poq/nfv048

51. *Barbera P. et al.* Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data / P. Barbera, A. Casas, J. Nagler [et al.]. // American Political Science Review. 2019. Vol. 113, No. 4. P. 883–901. DOI: 10.1017/S0003055419000352

52. *Девятко И.Ф.* Диагностическая процедура в социологии: Очерк истории и теории. М.: Наука, 1993. 175 с.

53. *Iannelli L. et al.* Facebook Digital Traces for Survey Research: Assessing the Efficiency and Effectiveness of a Facebook Ad–Based Procedure for Recruiting Online Survey Respondents in Niche and Difficult-to-Reach Populations / L. Iannelli, F. Giglietto, L. Rossi, E. Zurovac // Social Science Computer Review. 2020. Vol. 38, No. 4. P. 462–476. DOI: 10.1177/0894439318816638

54. *Kosinski M., Stillwell D., Graepel T*. Private traits and attributes are predictable from digital records of human behavior // Proceedings of the national academy of sciences. 2013. Vol. 110, No. 15. P. 5802–5805. DOI: 10.1073/pnas.1218772110

55. *Девятко И.Ф.* От «виртуальной лаборатории» до «социального телескопа»: метафоры тематических и методологических инноваций в онлайн-исследованиях // Онлайн-исследования в России: тенденции и перспективы / Под общ. ред. А.В. Шашкина, И.Ф. Девятко, С.Г. Давыдова. М.: Онлайн маркет интеллидженс, 2016. С. 19–33.

56. *Afriat H. et al.* “This is capitalism. It is not illegal”: Users’ attitudes toward institutional privacy following the Cambridge Analytica scandal / H. Afriat, S. Dvir-Gvirsman, K. Tsuriel, L. Ivan // The Information Society. 2021. Vol. 37, No. 2. P. 115–127. DOI: 10.1080/01972243.2020.1870596

57. *Diaz F. et al.* Online and Social Media Data As an Imperfect Continuous Panel Survey / F. Diaz, M. Gamon, J. M. Hofman, E. Kıcıman, D. Rothschild // PLoS ONE 2016. Vol. 11, No. 1. P. e0145406. DOI: 10.1371/journal.pone.0145406

58. *Бызов А.А*. Интеллектуальный анализ текстов в социальных науках // Социология: методология, методы, математическое моделирование (Социология: 4М). 2019. № 49. С. 131–160.

59. *Lazer D. et al.* The Parable of Google Flu: Traps in Big Data Analysis / D. Lazer, R. Kennedy, G. King, A. Vespignani // Science. 2014. Vol. 343, No. 6176. P. 1203–1205. DOI: 10.1126/science.1248506

60. *Hofstra B. et al. B.* Sources of Segregation in Social Networks: A Novel Approach Using Facebook / B. Hofstra, R. Corten, F. van Tubergen, N.B. Ellison // American Sociological Review. 2017. Vol. 82, No. 3. P. 625–656. DOI: 10.1177/0003122417705656

61. *Henderson M. et al.* Measuring Twitter Use: Validating Survey-Based Measures / M. Henderson, K. Jiang, M. Johnson, L. Porter // Social Science Computer Review. 2021. Vol. 39, No. 6. P. 1121–1141. DOI: 10.1177/0894439319896244

62. *Vraga E.K., Tully M.* Who Is Exposed to News? It Depends on How You Measure: Examining Self-Reported Versus Behavioral News Exposure Measures // Social Science Computer Review. 2020. Vol. 38, No. 5. P. 550–566. DOI: 10.1177/0894439318812050

63. *Haenschen K.* Self-Reported Versus Digitally Recorded: Measuring Political Activity on Facebook // Social Science Computer Review. 2020. Vol. 38, No. 5. P. 567–583. DOI: 10.1177/0894439318813586

64. *Jürgens P., Stark B., Magin M.* Two Half-Truths Make a Whole? On Bias in Self-Reports and Tracking Data // Social Science Computer Review. 2020. Vol. 38, No. 5. P. 600–615. DOI: 10.1177/0894439319831643

65. *Shin J.* How Do Partisans Consume News on Social Media? A Comparison of Self-Reports With Digital Trace Measures Among Twitter Users // Social Media + Society. 2020. Vol. 6, No. 4. DOI: 10.1177/2056305120981039

66. *Hopp T. et al.* Correlating Self-Report and Trace Data Measures of Incivility: A Proof of Concept / T. Hopp, C. J. Vargo, L. Dixon, N. Thain // Social Science Computer Review. 2020. Vol. 38, No. 5. P. 584–599. DOI: 10.1177/0894439318814241

67. *Junco R.* Comparing Actual and Self-Reported Measures of Facebook Use // Computers in Human Behavior. 2013. Vol. 29, No. 3. P. 626–631. DOI: 10.1016/j.chb.2012.11.007

68. *Hessler J.* Peoplemeter Technologies and the Biometric Turn in Audience Measurement // Television & New Media. 2021. Vol. 22, No. 4. P. 400–419. DOI: 10.1177/1527476419879415

69. *Parry D.A. et al.* A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use / D.A. Parry, B.I. Davidson, C.J.R. Sewall, J.T. Fisher, H. Mieczkowski, D.S. Quintana // Nature Human Behaviour. 2021. Vol. 5, No. 11. P. 1535–1547. DOI: 10.1038/s41562-021-01117-5

70. *Greenberg B.S.et al.* Comparing Survey and Diary Measures of Internet and Traditional Media Use / B.S. Greenberg, M.S. Eastin, P. Skalski, L. Cooper, M. Levy, K. Lachlan // Communication Reports. 2005. Vol. 18, No. 1–2. DOI: 10.1080/08934210500084164

71. *Araujo T. et al.* 'How Much Time Do You Spend Online? Understanding and Improving the Accuracy of Self-Reported Measures of Internet Use / T. Araujo, A. Wonneberger, P. Neijens, C. de Vreese // Communication Methods and Measures. 2017. Vol. 11, No. 3. P.173–90. DOI: 10.1002/9781119011071.iemp0049

72. *Wonneberger A., Irazoqui M.* Tell it like it is? Inaccuracies of self-reported TV exposure in comparison to people-meter data // Annual Conference of the International Communication Association. London, UK. 17–21 June 2013.

73. *Prior M.* The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure // Public Opinion Quarterly. 2009. Vol. 73, No. 1. P. 130–143. DOI: 10.1093/poq/nfp002

74. *Boase J., Ling R.* Measuring Mobile Phone Use: Self-Report Versus Log Data // Journal of Computer-Mediated Communication. 2013. Vol. 18, No. 4. P. 508–519. DOI: 10.1111/jcc4.12021

75. *Ettema J.S.* Explaining information system use with system-monitored vs. self-reported use measures // Public Opinion Quarterly. 1985. Vol. 49, No. 3. P. 381–387. DOI: 10.1086/268935

76. *van der Voort T.H.A., Vooijs M.W.* Validity of children's direct estimates of time spent television viewing // Journal of Broadcasting & Electronic Media. 1990. Vol. 34, No. 1. P. 93–99. DOI: 10.1080/08838159009386729

77. *Chang L.C., Krosnick J.A.* Measuring the frequency of regular behaviors: Comparing the "typical week" to the "past week" // Sociological Methodology. 2003. Vol. 33, No. 1. P. 55–80. DOI: 10.1111/j.0081-1750.2003.t01-1-00127.x

78. *Yanovitzky I.* Effect of Call-In Political Talk Radio Shows on Their Audiences: Evidence from a Multi-Wave Panel Analysis // International Journal of Public Opinion Research. 2001. Vol. 13, No. 4. P. 377–397. DOI: 10.1093/ijpor/13.4.377

79. *de Vreese C., Semetko H.A.* News matters: Influences on the vote in the Danish 2000 euro referendum campaign // European Journal of Political Research. 2004. Vol. 43, No. 5. P. 699–722. DOI: 10.1111/j.0304-4130.2004.00171.x

80. *van Spanje J., de Vreese C.* Europhile Media and Eurosceptic Voting: Effects of News Media Coverage on Eurosceptic Voting in the 2009 European Parliamentary Elections // Political Communication. 2014. Vol. 31, No. 2. P. 325–354. DOI: 10.1080/10584609.2013.828137

81. *McCombs M.E., Shaw D.L.* The Agenda-Setting Function of Mass Media // Public Opinion Quarterly. 1972. Vol. 36, No. 2. P. 176–187. DOI: 10.1086/267990

82. *Geers S., Bos L.* Priming Issues, Party Visibility, and Party Evaluations: The Impact on Vote Switching // Political Communication. 2017. Vol. 34, No. 3. P. 344–366. DOI: 10.1080/10584609.2016.1201179

83. *Hopmann D.N. et al.* Effects of Election News Coverage: How Visibility and Tone Influence Party Choice / D.N. Hopmann, R. Vliegenthart, C.H. De Vreese, E. Albæk // Political Communication. 2010. Vol. 27, No. 4. P. 389–405. DOI: 10.1080/10584609.2010.516798

84. *Matthes J.* Exposure to Counterattitudinal News Coverage and the Timing of Voting Decisions // Communication Research. 2012. Vol. 39, No. 2. P. 147–169. DOI: 10.1177/0093650211402322

85. *Mellon J., Prosser C.* Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users // Research & Politics. 2017. Vol. 4, No. 3. DOI: 10.1177/2053168017720008

86. *Stier S. et al.* Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter / S. Stier, A. Bleier, H. Lietz, M. Strohmaier // Political Communication. 2018. Vol. 35, No. 1. P. 50–74. DOI: 10.1080/10584609.2017.1334728

87. *Beauchamp N.* Predicting and Interpolating State-Level Polls Using Twitter Textual Data // American Journal of Political Science. 2017. Vol. 61, No. 2. P. 490–503. DOI: 10.1111/ajps.12274

88. *O'Connor B. et al.* From tweets to polls: Linking text sentiment to public opinion time series / B. O'Connor, R. Balasubramanyan, B. R. Routledge, N.A. Smith // Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. Palo Alto, CA: AAAI Press, 2010. P. 122–129.

89. *Olsson T.* An indispensable resource: The Internet and young civic engagement // Young Citizens and New Media: Learning for democratic participation. New York: Routledge, 2013. P. 197–214.

90. *Bennett W.L., Wells C., Freelon D.* Communicating citizenship online: Models of civic learning in the youth web sphere // A Report from the Civic Learning Online Project. 2009. 41 p.

91. *Giglietto F.* If Likes Were Votes: An Empirical Study on the 2011 Italian Administrative Elections // SSRN Journal. 7 May 2012. DOI: 10.2139/ssrn.1982736

92. *MacWilliams M.C.* Forecasting Congressional Elections Using Facebook Data // APSC. 2015. Vol. 48, No. 4. P. 579–583. DOI: 10.1017/S1049096515000797

93. *DiGrazia J. et al.* More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior / J. DiGrazia, K. McKelvey, J. Bollen, F. Rojas // PLoS ONE. 2013. Vol. 8, No. 11. DOI: 10.1371/journal.pone.0079449

94. *Bach R.L. et al.* Predicting Voting Behavior Using Digital Trace Data / R.L. Bach, C. Kern, A. Amaya, F. Keusch, F. Kreuter, J. Hecht, J. Heinemann // Social Science Computer Review. 2021. Vol. 39, No. 5. P. 862–883. DOI: 10.1177/0894439319882896

95. *Jungherr A., Jürgens P., Schoen H.* Why the Pirate Party Won the German Election of 2009 or The Trouble with Predictions: A Response to Tumasjan A., Sprenger T.O., Sander P.G., & Welpe I.M. "Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment" // Social Science Computer Review. 2012. Vol. 30, No. 2. P. 229–234. DOI: 10.1177/0894439311404119

96. *Guess A.M.* Measure for Measure: An Experimental Test of Online Political Media Exposure // Political Analysis. 2015. Vol. 23, No. 1. P. 59–75. DOI: 10.1093/pan/mpu010

97. *Vraga E., Bode L., Troller-Renfree S.* Beyond Self-Reports: Using Eye Tracking to Measure Topic and Style Differences in Attention to Social Media Content //

Communication Methods and Measures. 2016. Vol. 10, No. 2–3. P. 149–164. DOI: 10.1080/19312458.2016.1150443

98. *Colleoni E., Rozza A., Arvidsson A.* Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data: Political Homophily on Twitter // Journal of Communication. 2014. Vol. 64, No. 2. P. 317–332. DOI: 10.1111/jcom.12084

99. *Nelson J.L., Webster J.G.* The Myth of Partisan Selective Exposure: A Portrait of the Online Political News Audience // Social Media + Society. 2017. Vol. 3, No. 3. DOI: 10.1177/2056305117729314

100. *Connelly R. et al.* The role of administrative data in the big data revolution in social science research / R. Connelly, C. Playford, V. Gayle, C. Dibben // Social Science Research. 2016. Vol. 59. DOI: 10.1016/j.ssresearch.2016.04.015

101. *Yoshida Y., Haan M., Schaffer S.* Administrative data linkage in Canada: Implications for sociological research // Canadian Review of Sociology. 2022. Vol. 59, No. 2. P. 251–270. DOI: 10.1111/cars.12376

102. *Harron K. et al.* Challenges in administrative data linkage for research / K. Harron, C. Dibben, J. Boyd, A. Hjern, M. Azimaee, M.L. Barreto, H. Goldstein // Big Data & Society. 2017. Vol. 4, No. 2. DOI: 10.1177/2053951717745678

103. *Choi K.H., Ramaj S., Haan M.* Age of the oldest child and internal migration of immigrant families: A study using administrative data from immigrant landing and tax files // Population Space and Place. 2021. Vol. 27, No. 4. DOI: 10.1002/psp.2409

104. *Rampazzo F. et al.* A Framework for Estimating Migrant Stocks Using Digital Traces and Survey Data: An Application in the United Kingdom / F. Rampazzo, J. Bijak, A. Vitali, I. Weber, E. Zagheni // Demography. 2021. Vol. 58, No. 6. P. 2193–2218. DOI: 10.1215/00703370-9578562

105. *Brown J.R. et al.* Childhood cross-ethnic exposure predicts political behavior seven decades later: Evidence from linked administrative data / J.R. Brown, R.D. Enos, J. Feigenbaum, S. Mazumder // Science Advances. 2021. Vol. 7, No. 24. DOI: 10.1126/sciadv.abe8432

106. *Vatsalan D. et al.* Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges / D. Vatsalan, Z. Sehili, P. Christen, E. Rahm // Handbook of Big Data Technologies / Ed. by A.Y. Zomaya, S. Sakr. Cham: Springer International Publishing, 2017. P. 851–895. DOI: 10.1007/978-3-319-49340-4_25

107. *Dibben C. et al.* The data linkage environment / C. Dibben, M. Elliot, H. Gowans, D. Lightfoot // Methodological Developments in Data Linkage. Chapter 3 / Ed. by K. Harron, C. Dibben, H. Goldstein. London: Wiley, 2015. P. 36–62. DOI: 10.1002/9781119072454.ch3

Saponova Anastasia V.,
Lecturer, Postgraduate student, Department of Social Institutions Analysis, HSE University, Moscow, Russia, asaponova@hse.ru

Kulikov Sergey P.,
Postgraduate student, Department of Social Institutions Analysis, HSE University, Moscow, Russia, spkulikov@hse.ru

Integration of survey data and digital footprints: an overview of the main methodological approaches

The main purpose of current study is to review the main existing methodological approaches to the integration of survey data and digital traces that are used in sociological research. The paper examines key arguments in the current methodological discussion about the place of big digital data in contemporary social science research. The authors make an attempt to scrutinize the practice of integrating survey data and digital traces through the concept of "reactive – nonreactive" measurement. The possible functions of digital traces in the design of the study are indicated (on the example of social media data). On the example of three research areas (the study of media consumption, media effects and electoral behavior) general methodological principles for integrating data of different nature are demonstrated and possible prospects for the development of these approaches is described. The article discusses a wide range of methodological issues: problems of the data linking validity; potential threats to the validity of digital traces; opportunities to improve survey questionnaire, to enrich data, to search for new valid indicators of socio-political processes and to provide cross-validation of research results. The current practices of integrating administrative data are considered as well.

Keywords: data integration, data linkage, big data, nonreactive research, digital traces, survey data

References

1. Golder S.A., Macy M.W. Digital Footprints: Opportunities and Challenges for Online Social Research, *Annual Review of Sociology*, 2014, 40 (1), 129–152. DOI: 10.1146/annurev-soc-071913-043145

2. Salganik M.J., Watts D.J. Web-Based Experiments for the Study of Collective Social Dynamics in Cultural Markets, *Topics in Cognitive Science*, 2009, 1 (3), 439–468. DOI: 10.1111/j.1756-8765.2009.01030.x
3. Anderson C. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, *Wired*, 23.06.2008. URL: https://www.wired.com/2008/06/pb-theory/ (access date: 17.04.2022).
4. Mayer-Schönberger V., Cukier K. *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (in Russian). Moscow: Mann, Ivanov, Ferber, 2014.
5. Boyd d., Crawford K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon, *Information, communication & society*, 2012, 15 (5), 662–679. DOI: 10.1080/1369118X.2012.678878
6. Guba K. Big Data in Sociology: New Data, New Sociology? (in Russian), *Sotsiologicheskoye obozreniye* (*Russian Sociological Review*), 2018, 17 (1), 213–236.
7. McFarland D. A., Lewis K., Goldberg A. Sociology in the era of big data: The ascent of forensic social science, *The American Sociologist*, 2016, 47 (1), 12–35. DOI: 10.1007/s12108-015-9291-8
8. Burrows R., Savage M. After the crisis? Big Data and the methodological challenges of empirical sociology, *Big Data & Society*, 2014, 1 (1). DOI: 10.1177/0038038507080443
9. Thrift N. *Knowing Capitalism*. London: SAGE Publications Ltd, 2005, 256 p. DOI: 10.4135/9781446211458
10. Gane N. Measure, Value and the Current Crises of Sociology, *The Sociological Review*, 2011, 59 (2), 151–173. DOI: 10.1111/j.1467-954X.2012.02054.x
11. Ignatow G. *Sociological Theory in the Digital Age*, 1st ed. New York: Routledge, 2020, 120 p. DOI: 10.4324/9780429292804
12. Kitchin R., Lauriault T.P. Small data in the era of big data, *GeoJournal*, 2015, 80 (4), 463–475. DOI: 10.1007/s10708-014-9601-7
13. de Vreese C.H., Boukes M., Schuck A., Vliegenthart R., Bos L., Lelkes Y. Linking Survey and Media Content Data: Opportunities, Considerations, and Pitfalls, *Communication Methods and Measures*, 2017, 11 (4), 221–244. DOI: 10.1080/19312458.2017.1380175
14. Stier S., Breuer J., Siegers P., Thorson K. Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field,

Social Science Computer Review, 2020, 38 (5), p. 503–516. DOI: 10.1177/0894439319843669

15. Beninger K., Digby A., Dillon G., MacGregor J. (eds.) Understanding Society: How people decide whether to give consent to link their administrative and survey data. In: *Understanding Society Working Paper Series*, 2017, no. 13, 65 p.
16. Webb E.J., Campbell D.T., Schwartz R.D., Sechrest L. *Unobtrusive measures: nonreactive research in the social sciences*, Chicago: Rand McNally, 1966, 220 p.
17. Bouchard Jr T.J. Unobtrusive Measures: An Inventory of Uses, *Sociological Methods & Research*, 1976, 4 (3), 267–300. DOI: 10.1177/004912417600400301
18. Hill A.D., White M.A., Wallace J.C. Unobtrusive measurement of psychological constructs in organizational research, *Organizational Psychology Review*, 2014, 4 (2), 148–174. DOI: 10.1177/2041386613505613
19. Lee R.M. Unobtrusive Methods, in: *Handbook of Research Methods in Health Social Sciences*. Ed. by P. Liamputtong. Wiesbaden: Springer VS, 2019, p. 491–507. DOI: 10.1007/978-981-10-5251-4_85
20. Deviatko I. Online research toolkit: an attempt at cataloging (in Russian), in: *Onlajn issledovaniya v Rossii 3.0*. Moskva: Online Market Intelligence, 2012, p. 17–31.
21. Dudina V. Digital Data as the Potential for the Development of Sociological Knowledge (in Russian), *Sotsiologicheskie issledovaniya (Sociological Studies*), 2016, 9, 21–30.
22. Lee R.M. *Unobtrusive Measures in Social Research*. Philadelphia, PA: Open University Press, 2000, 192 p.
23. Kalokyri V., Borgida A., Marian A., Vianna D. Integration and Exploration of Connected Personal Digital Traces, in: *Proceedings of the ExploreDB'17*. Chicago, IL: ACM, 2017. DOI: 10.1145/3077331.3077337
24. Araujo T., Neijens P. Unobtrusive Measures for Media Research, in: *The International Encyclopedia of Media Psychology*, 1st ed. Ed. by J. Bulck. Hoboken. NJ: Wiley Blackwell, 2020, p. 1–7. DOI: 10.1002/9781119011071.iemp0049
25. Deviatko I. New Data, New Statistics: from Reproducibility Crisis Toward New Requirements to Data Analysis and Presentation in Social

Sciences (in Russian), *Sotsiologicheskie issledovaniya (Sociological Studies)*, 2018, 12, 30–38.
26. Nikolaenko G., Fedorova A. Non-Reactive Strategy: Unobtrusive Methods of Gathering Sociological Information in Web 2.0 Age − Evidence from Digital Ethnography and Big Data (in Russian), *Sociologiya vlasti (Sociology of Power)*, 2017, 9 (4), 36–54.
27. Savage M., Burrows R. The Coming Crisis of Empirical Sociology, *Sociology*, 2007, 41 (5), 885–899. DOI: 10.1177/0038038507080443
28. Couper M.P. Is the Sky Falling? New Technology, Changing Media, and the Future of Survey", *Survey Research Methods*. European Survey Research Association, 2013, 7 (3), 145–156. DOI: 10.18148/SRM/2013.V7I3.5751
29. Beaulieu A. Sociable hyperlinks: an ethnographic approach to connectivity, in: *Virtual Methods: issues in social research on the Internet*. Ed. by C. Hine. Oxford: Berg, 2005, p. 183–198.
30. Hine C. Internet Research and Unobtrusive Methods, *Social Research Update*, 2011, 61, 1–4.
31. Dirksen V., Huizing A., Smit B. 'Piling on layers of understanding': the use of connective ethnography for the study of (online) work practices, *New Media & Society*, 2010, 12 (7), 1045–1063. DOI: 10.1177/1461444809341437
32. De Heer W., De Leeuw E. Trends in household survey nonresponse: A longitudinal and international comparison, *Survey nonresponse*, 2002, 41, 41–54.
33. Tourangeau R., Plewes T.J. (eds.) *Nonresponse in Social Science Surveys: A Research Agenda*. Washington, DC: The National Academies Press, 2013, 151 p.
34. Čehovin G., Bosnjak M., Lozar Manfreda K. Item Nonresponse in Web Versus Other Survey Modes: A Systematic Review and Meta-Analysis, *Social Science Computer Review*, 2022. DOI: 10.1177/08944393211056229
35. Lazer D., Pentland A., Adamic L., Aral S., Barabási A.-L., Brewer D., Christakis N., Contractor N., Fowler J., Gutmann M., Jebara T., King G., Macy M., Roy D., Van Alstyne M. Computational social science, *Science*, 2009, 323 (5915), 721-723. DOI: 10.1126/science.1167742
36. Keeter S., Christian L. *A comparison of results from surveys by the Pew Research Center and Google Consumer Surveys*. Washington, DC: Pew Research Center, 2012, 30 p.

37. Graham M., Hale S. A., Gaffney D. Where in the world are you? Geolocation and language identification in Twitter, *The Professional Geographer*, 2014, 66 (4), 568–578. DOI: 10.1080/00330124.2014.907699
38. Conrad F.G., Gagnon-Bartsch J. A., Ferg, R. A., Schober M. F., Pasek J., Hou E. Social Media as an Alternative to Surveys of Opinions About the Economy, *Social Science Computer Review*, 2021, 39 (4), 489–508. DOI: 10.1177/0894439319875692
39. Schulz A., Hadjakos A., Paulheim H., Nachtwey J., & Mühlhäuser M. A Multi-Indicator Approach for Geolocalization of Tweets, *ICWSM*, 2021, 7 (1), 573–582.
40. Stock K. Mining location from social media: A systematic review, *Computers, Environment and Urban Systems*, 2018, 71, 209–240. DOI: 10.1016/j.compenvurbsys.2018.05.007
41. Chen Q., Poorthuis A. Identifying home locations in human mobility data: an open-source R package for comparison and reproducibility, *International Journal of Geographical Information Science*, 2021, 35 (7), 1425–1448. DOI: 10.1080/13658816.2021.1887489
42. Campbell D.T., Fiske D.W. Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological bulletin*, 1959, 56 (2), 81–105. DOI: 10.1037/h0046016
43. Bouchard Jr T.J. Field research methods: Interviewing, questionnaires, participant observation, systematic observation, unobtrusive measures, in: *Handbook of industrial and organizational psychology*, vol. 1. Ed. by M.D. Dunnette. Chicago: Rand McNally, 1976, p. 363–413.
44. Zeller R.A., Carmines E.G. *Measurement in the social sciences: the link between theory and data*. Cambridge; New York: Cambridge University Press, 1980, 198 p.
45. Pasek J., McClain C. A., Newport F., Marken S. Who's Tweeting About the President? What Big Survey Data Can Tell Us About Digital Traces?, *Social Science Computer Review*, 2020, 38 (5), 633–650. DOI: 10.1177/0894439318822007
46. Klimova A. M., Kulikov S. P., Chmel K. S. The Role of Social Media in Shaping Regional Ecological Protest in Russia (in Russian), *Monitoring of Public Opinion: Economic and Social Changes*, 2021, 6 (28), 28–52. DOI: 10.14515/monitoring.2021.6.2024

47. Shlomo N. Overview of Data Linkage Methods for Policy Design and Evaluation, in: *Data-Driven Policy Impact Evaluation*. Ed. by N. Crato, P. Paruolo. Cham: Springer International Publishing, 2019, p. 47–65.
48. Quinlan S., Gummer T., Roßmann J., Wolf C. 'Show me the money and the party!' – variation in Facebook and Twitter adoption by politicians, *Information, Communication & Society*, 2018, 21 (8), 1031–1049. DOI: 10.1080/1369118X.2017.1301521
49. Karlsen R., Enjolras B. Styles of Social Media Campaigning and Influence in a Hybrid Political Communication System: Linking Candidate Survey Data with Twitter Data, *The International Journal of Press/Politics*, 2016, 21 (3), 338–357. DOI: 10.1177/1940161216645335
50. Schober M.F., Pasek J., Guggenheim L. Social Media Analyses for Social Measurement. *Public Opinion Quarterly*, 2016, 80 (1), 180–211. DOI: 10.1093/poq/nfv048
51. Barbera P., Casas A., Nagler J. "Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data", *American Political Science Review*, 2019, 113 (4), 883–901. DOI: 10.1017/S0003055419000352
52. Deviatko I. *Diagnostic procedure in sociology* (in Russian). Moscow: Nauka, 1993.
53. Iannelli L., Giglietto F., Rossi L., Zurovac E. Facebook Digital Traces for Survey Research: Assessing the Efficiency and Effectiveness of a Facebook Ad–Based Procedure for Recruiting Online Survey Respondents in Niche and Difficult-to-Reach Populations, *Social Science Computer Review*, 2020, 38 (4), 462–476. DOI: 10.1177/0894439318816638
54. Kosinski M., Stillwell D., Graepel T. Private traits and attributes are predictable from digital records of human behavior, *Proceedings of the national academy of sciences*, 2013, 110 (15), 5802–5805. DOI: 10.1073/pnas.1218772110
55. Deviatko I. From "Virtual Lab" to "Social Telescope": Metaphors of Theoretical and Methodological Innovations in Online Research (in Russian), in: *Onlajn issledovaniya v Rossii 4.0.* Moskva: Online Market Intelldgence, 2016, p. 19–33.
56. Afriat H., Dvir-Gvirsman S., Tsuriel K., Ivan L. "This is capitalism. It is not illegal": Users' attitudes toward institutional privacy following the Cambridge Analytica scandal, *The Information Society*, 2021, 37 (2), 115–127. DOI: 10.1080/01972243.2020.1870596

57. Diaz F., Gamon M., Hofman J. M., Kıcıman E., Rothschild D. Online and Social Media Data As an Imperfect Continuous Panel Survey, *PLoS ONE*, 2016, 11 (1), e0145406. DOI: 10.1371/journal.pone.0145406
58. Byzov A. Text mining in social sciences (in Russian), *Sotsiologiya 4M (Sociology: methodology, methods, mathematical modeling)*, 2019, 49, 131–160.
59. Lazer D., Kennedy R., King G., Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis, *Science*, 2014, 343 (6176), 1203–1205. DOI: 10.1126/science.1248506
60. Hofstra B., Corten R., van Tubergen F., Ellison N. B. Sources of Segregation in Social Networks: A Novel Approach Using Facebook, *American Sociological Review*, 2017, 82 (3), 625–656. DOI: 10.1177/0003122417705656
61. Henderson M., Jiang K., Johnson M., Porter L. Measuring Twitter Use: Validating Survey-Based Measures, *Social Science Computer Review*, 2021, 39 (6), 1121–1141. DOI: 10.1177/0894439319896244
62. Vraga E.K., Tully M. Who Is Exposed to News? It Depends on How You Measure: Examining Self-Reported Versus Behavioral News Exposure Measures, *Social Science Computer Review*, 2020, 38 (5), 550–566. DOI: 10.1177/0894439318812050
63. Haenschen K. Self-Reported Versus Digitally Recorded: Measuring Political Activity on Facebook, *Social Science Computer Review*, 2020, 38 (5), 567–583. DOI: 10.1177/0894439318813586
64. Jürgens P., Stark B., Magin M. Two Half-Truths Make a Whole? On Bias in Self-Reports and Tracking Data, *Social Science Computer Review*, 2020, 38 (5), 600–615. DOI: 10.1177/0894439319831643
65. Shin J. How Do Partisans Consume News on Social Media? A Comparison of Self-Reports With Digital Trace Measures Among Twitter Users, *Social Media + Society*, 2020, 6 (4). DOI: 10.1177/2056305120981039
66. Hopp T., Vargo C. J., Dixon L., Thain N. Correlating Self-Report and Trace Data Measures of Incivility: A Proof of Concept, *Social Science Computer Review*, 2020, 38 (5), 584–599. DOI: 10.1177/0894439318814241
67. Junco R. Comparing Actual and Self-Reported Measures of Facebook Use, *Computers in Human Behavior*, 2013, 29 (3), 626–631. DOI: 10.1016/j.chb.2012.11.007

68. Hessler J. Peoplemeter Technologies and the Biometric Turn in Audience Measurement, *Television & New Media*, 2021, 22 (4), 400–419. DOI: 10.1177/1527476419879415
69. Parry D.A., Davidson B. I., Sewall C. J. R., Fisher J. T., Mieczkowski H., Quintana, D. S. A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use, *Nature Human Behaviour*, 2021, 5 (11), 1535–1547. DOI: 10.1038/s41562-021-01117-5
70. Greenberg B.S., Eastin M. S., Skalski P., Cooper L., Levy M., Lachlan K. Comparing Survey and Diary Measures of Internet and Traditional Media Use, *Communication Reports*, 2005, 18 (1–2). DOI: 10.1080/08934210500084164
71. Araujo, T., Wonneberger A., Neijens P., de Vreese C.H. How Much Time do You Spend Online? Understanding and Improving the Accuracy of Self-Reported Measures of Internet Use, *Communication Methods and Measures*, 2017, 11 (3), 173–190. https://doi.org/10.1080/19312458.2017.1317337
72. Wonneberger A., Irazoqui M. Tell it like it is? Inaccuracies of self-reported TV exposure in comparison to people-meter data, *Annual Conference of the International Communication Association*. London, UK. 17–21 June 2013.
73. Prior M. The Immensely Inflated News Audience: Assessing Bias in Self-Reported News Exposure, *Public Opinion Quarterly*, 2009, 73 (1), 130–143. DOI: 10.1093/poq/nfp002
74. Boase J., Ling R. Measuring Mobile Phone Use: Self-Report Versus Log Data, *Journal of Computer-Mediated Communication*, 2013, 18 (4), 508–519. DOI: 10.1111/jcc4.12021
75. Ettema J.S. Explaining information system use with system-monitored vs. self-reported use measures, *Public Opinion Quarterly*, 1985, 49 (3), 381–387. DOI: 10.1086/268935
76. van der Voort T.H.A., Vooijs M.W. Validity of children's direct estimates of time spent television viewing, *Journal of Broadcasting & Electronic Media*, 1990, 34 (1), 93–99. DOI: 10.1080/08838159009386729
77. Chang L.C., Krosnick J.A. Measuring the frequency of regular behaviors: Comparing the "typical week" to the "past week", *Sociological Methodology*, 2003, 33 (1), 55–80. DOI: 10.1111/j.0081-1750.2003.t01-1-00127.x

78. Yanovitzky I. Effect of Call-In Political Talk Radio Shows on Their Audiences: Evidence from a Multi-Wave Panel Analysis, *International Journal of Public Opinion Research*, 2001, 13 (4), 377–397. DOI: 10.1093/ijpor/13.4.377
79. de Vreese C.H., Semetko H.A. News matters: Influences on the vote in the Danish 2000 euro referendum campaign, *European Journal of Political Research*, 2004, 43 (5), 699–722. DOI: 10.1111/j.0304-4130.2004.00171.x
80. van Spanje J., de Vreese C. Europhile Media and Eurosceptic Voting: Effects of News Media Coverage on Eurosceptic Voting in the 2009 European Parliamentary Elections, *Political Communication*, 2014, 31 (2), 325–354. DOI: 10.1080/10584609.2013.828137
81. McCombs M.E., Shaw D.L. The Agenda-Setting Function of Mass Media, *Public Opinion Quarterly*, 1972, 36 (2), 176–187. DOI: 10.1086/267990
82. Geers S., Bos L. Priming Issues, Party Visibility, and Party Evaluations: The Impact on Vote Switching, *Political Communication*, 2017, 34 (3), 344–366. DOI: 10.1080/10584609.2016.1201179
83. Hopmann D.N., Vliegenthart R., De Vreese C. H., Albæk E. Effects of Election News Coverage: How Visibility and Tone Influence Party Choice, *Political Communication*, 2010, 27 (4), 389–405. DOI: 10.1080/10584609.2010.516798
84. Matthes J. Exposure to Counterattitudinal News Coverage and the Timing of Voting Decisions, *Communication Research*, 2012, 39 (2), 147–169. DOI: 10.1177/0093650211402322
85. Mellon J., Prosser C. Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users, *Research & Politics*, 2017, 4 (3). DOI: 10.1177/2053168017720008
86. Stier S., Bleier A., Lietz H., Strohmaier M. Election Campaigning on Social Media: Politicians, Audiences, and the Mediation of Political Communication on Facebook and Twitter, *Political Communication*, 2018, 35 (1), 50–74. DOI: 10.1080/10584609.2017.1334728
87. Beauchamp N. Predicting and Interpolating State-Level Polls Using Twitter Textual Data, *American Journal of Political Science*, 2017, 61 (2), 490–503. DOI: 10.1111/ajps.12274

88. O'Connor B., Balasubramanyan R., Routledge B.R., Smith, N.A. From tweets to polls: Linking text sentiment to public opinion time series, in: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. Palo Alto, CA: AAAI Press, 2010, p. 122–129.
89. Olsson T. An indispensable resource: The Internet and young civic engagement, in: *Young Citizens and New Media: Learning for democratic participation*. New York: Routledge, 2013, p. 197–214.
90. Bennett W.L., Wells C., Freelon D. Communicating citizenship online: Models of civic learning in the youth web sphere, in: *A Report from the Civic Learning Online Project*, 2009, 41 p.
91. Giglietto F. If Likes Were Votes: An Empirical Study on the 2011 Italian Administrative Elections, *SSRN Journal*, 7 May 2012. DOI: 10.2139/ssrn.1982736
92. MacWilliams M.C. Forecasting Congressional Elections Using Facebook Data, *APSC*, 2015, 48 (4), 579–583. DOI: 10.1017/S1049096515000797
93. DiGrazia J., McKelvey K., Bollen J., Rojas F. More Tweets, More Votes: Social Media as a Quantitative Indicator of Political Behavior, *PLoS ONE*, 2013, 8 (11). DOI: 10.1371/journal.pone.0079449
94. Bach R.L., Kern C., Amaya A., Keusch F., Kreuter F., Hecht J., Heinemann J. Predicting Voting Behavior Using Digital Trace Data, *Social Science Computer Review*, 2021, 39 (5), 862–883. DOI: 10.1177/0894439319882896
95. Jungherr A., Jürgens P., Schoen H. Why the Pirate Party Won the German Election of 2009 or The Trouble with Predictions: A Response to Tumasjan A., Sprenger T.O., Sander P.G., & Welpe I.M. "Predicting Elections with Twitter: What 140 Characters Reveal About Political Sentiment", *Social Science Computer Review*, 2012, 30 (2), 229–234. DOI: 10.1177/0894439311404119
96. Guess A.M. Measure for Measure: An Experimental Test of Online Political Media Exposure, *Political Analysis*, 2015, 23 (1), 59–75. DOI: 10.1093/pan/mpu010
97. Vraga E., Bode L., Troller-Renfree S. Beyond Self-Reports: Using Eye Tracking to Measure Topic and Style Differences in Attention to Social Media Content, *Communication Methods and Measures*, 2016, 10 (2–3), 149–164. DOI: 10.1080/19312458.2016.1150443

98. Colleoni E., Rozza A., Arvidsson A. Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data: Political Homophily on Twitter, *Journal of Communication*, 2014, 64 (2), 317–332. DOI: 10.1111/jcom.12084
99. Nelson J.L., Webster J.G. The Myth of Partisan Selective Exposure: A Portrait of the Online Political News Audience, *Social Media + Society*, 2017, 3 (3), 1–13. DOI: 10.1177/2056305117729314
100. Connelly R., Playford C., Gayle V., Dibben C. The role of administrative data in the big data revolution in social science research, *Social Science Research*, 2016, 59, 1–12. DOI: 10.1016/j.ssresearch.2016.04.015
101. Yoshida Y., Haan M., Schaffer S. Administrative data linkage in Canada: Implications for sociological research, *Canadian Review of Sociology*, 2022, 59 (2), 251–270. DOI: 10.1111/cars.12376
102. Harron K., Dibben C., Boyd J., Hjern A., Azimaee M., Barreto M.L., Goldstein H. Challenges in administrative data linkage for research, *Big Data & Society*, 2017, 4 (2). DOI: 10.1177/2053951717745678
103. Choi K.H., Ramaj S., Haan M. Age of the oldest child and internal migration of immigrant families: A study using administrative data from immigrant landing and tax files, *Population Space and Place*, 2021, 27 (4). 4. DOI: 10.1002/psp.2409
104. Rampazzo F., Bijak J., Vitali A., Weber I., Zagheni, E. A Framework for Estimating Migrant Stocks Using Digital Traces and Survey Data: An Application in the United Kingdom, *Demography*, 2021, 58 (6), 2193–2218. DOI: 10.1215/00703370-9578562
105. Brown J.R., Enos R.D., Feigenbaum J., Mazumder S. Childhood cross-ethnic exposure predicts political behavior seven decades later: Evidence from linked administrative data, *Science Advances*, 2021, 7 (24). DOI: 10.1126/sciadv.abe8432
106. Vatsalan D., Sehili, Z., Christen, P., Rahm, E. Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges, in: *Handbook of Big Data Technologies*. Ed. by A.Y. Zomaya, S. Sakr. Cham: Springer International Publishing, 2017, p. 851–895. DOI: 10.1007/978-3-319-49340-4_25
107. Dibben C., Elliot M., Gowans H., Lightfoot D. The data linkage environment, in: *Methodological Developments in Data Linkage*. Chapter 3. London: Wiley, 2015, p. 36–62. DOI: 10.1002/9781119072454.ch3