УДК 303.688
DOI: 10.17223/1998863X/54/16

О.В. Вилкова

К ВОПРОСУ О НАУЧНОЙ ОСМЫСЛЕННОСТИ ПРИМЕНЕНИЯ ВЕБ-СКРЕЙПИНГА КАК МЕТОДА СБОРА ДАННЫХ В СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЯХ

Статья посвящена современному методу сбора открытых интернет-данных – веб-скрейпингу – и научной осмысленности его использования в социологических исследованиях. Основываясь на концепциях цифровой социологии, приводятся методологические и технические возможности и ограничения веб-скрейпинга. С позиции философии науки обосновывается место, отводимое веб-скрейпингу в структуре социологического знания. Приведен обзор недавних социологических исследований с применением данного метода.

Ключевые слова: *веб-скрейпинг, цифровая социология, социология интернета, методы социологических исследований.*

Введение

С ростом цифровизации и постиндустриальными изменениями в социологической науке выделилось отдельное направление – цифровая социология. В 2015 г. на первой в мире конференции по цифровой социологии этой предметной области было дано определение субдисциплины, которая изучает цифровые средства, платформы и технологии как элементы повседневности, а также их влияние на поведение человека в группе – с использованием методов извлечения данных цифровых платформ [1. P. 1]. Сегодня социологи вынуждены применять новые подходы к изучению трансформирующихся социальных реалий. Объектом исследования становится не индивид, а «социотехнический гибрид» (пользователь приложения, вебсайта, гаджета) и его «методологический аккаунт»; гаджеты накапливают информацию о социальном поведении, доступную для анализа, – так называемые большие социологические данные [2. С. 21].

Цифровая социология активно развивается, чувствительно реагируя на социально-экономические и политические изменения. В 2017 г. после голосования по Brexit и провозглашения нового президента США Д. Трампа выходит книга Нортье Маррес, значимого теоретика цифровой социологии, «Digital Sociology: The Reinvention of Social Research». В книге Маррес анализирует релевантность методов анализа цифровых данных для решения социологических задач. Книга оказалась востребованной, поскольку была издана после событий, породивших дискуссии о роли социальных сетей, медиа, фейковых новостях и инструментах их идентификации. Проявило интерес к цифровой социологии и отечественное академическое сообщество: в конце лета 2018 г. в Государственном университете управления учреждается журнал «Цифровая социология».

Несмотря на признание академической актуальности, цифровая социология подвергается критике в отношении методов сбора и анализа данных.

Д. Фаррелл и Дж. Петерсен, считавшиеся пионерами веб-технологий в социологии и призывавшие к сбору и анализу неструктурированных интернет-данных [3. P. 1], в последних работах используют интернет для сбора структурированной информации – отдавая предпочтение интернет-опросам [4. P. 1]. Анализ интернет-данных ученые считают проблематичным в силу возможного нарушения приватности данных и несоответствия показателям теоретической рамки исследования [5. P. 2] (табл. 1).

Таблица 1. **Критические положения несоответствия измерений, проводимых методами цифровой социологии, теоретической рамке исследования**

Положение		Влияние на исследование
1. Нерепрезентативность выборок интернет-данных		
1.1.	Невозможно доказать репрезентативность: гарантировать, что данная выборка пользователей случайная и подчинена закону больших чисел	Описывается частный случай, а не массовый социальный процесс / группа
1.2.	Анонимность и безнаказанность интернет-среды искажет выборку за счёт спама, бот-активности и недостоверной информации	Состояние социальной группы / процесса искажается за счёт нерелевантных данных
2. Интернет как площадка искажает дискурс		
2.1.	При оценке общественных мнений / настроений порог входа в интернет как площадку низкий; культура интернет-среды заменяет рациональные аргументы жёсткими эпитетами и обвинениями	Описывается не действительное состояние социальной группы / процесса, а семантические искажения, связанные с общедоступностью интернет-среды
2.2.	Несмотря на всеобщую цифровизацию, имеет место цифровое неравенство: отдельные социальные группы могут быть не охвачены средствами, оставляющими цифровой след; существует проблема доступа к отдельным порталам; технические неполадки в работе интернет-платформы и т.п.	Охватывается не вся выборка объекта / предмета исследования из теоретической рамки исследования
2.3.	Постиндустриальное общество необходимо рассматривать не только как пространство онлайн-среды, но и как офлайн-среды [6. С. 31]. Сложность при таком анализе заключается в том, что не всегда онлайн- и офлайн-статусы индивида совпадают; необходимо искать пересечения между онлайн- и офлайн-статусом индивида, а не выстраивать их как автономные структуры	Рассматривается только одна сторона медали - виртуальная действительность индивида; теоретическая рамка может включать и его офлайн-статус, который с помощью цифровых данных сложно описать
2.4.	Цифровая социология стремится проводить анализ в режиме real-time, опираясь на потоки сообщений блогосферы, RSS-ленты, на лету формируя выборки и разрезы, отсеивая аномалии и выбросы [7. С. 6]	Real-time извлечение и фильтрация данных могут дать существенные искажения на более длинном горизонте анализа, если таковая определялась теоретической рамкой исследования
3. Нарушение приватности персональных данных: сбор информации должен сопровождаться согласием носителей данных		Угроза правонарушения и наказания для исследователя
4. Увлечение математическими методами: социологи в погоне за оптимальной математической спецификацией забывают основную задачу – описать социальную группу, институт или процесс [8. P. 32]		Цели и задачи теоретической рамки, замысел исследования не раскрываются

Социологические исследования используют современные стратегии анализа данных, опираясь на теоретические описания их преимуществ [9. С. 132; 10. С. 4]. Однако исследования, системно описывающие возможности и ограничения метода сбора данных цифровой социологии – веб-скрейпинга – и его научной осмысленности, отсутствуют. В условиях тренда на доказательную социальную науку, призывающую «идти от данных» [11. С. 217], применение веб-скрейпинга актуализируется. Работа ставит задачей систематизировать возможности и ограничения веб-скрейпинга для преодоления стигматизации вокруг сбора интернет-данных в социологических исследованиях.

Веб-скрейпинг в социологических исследованиях: определение, возможности, ограничения

Веб-скрейпинг (син. парсинг, скрин-скрейпинг, веб-кроулинг) – практика сбора открытых данных с загруженных веб-страниц и форм, не предназначенных для этого (т.е. в большинстве случаев в обход интерфейсных правил пользования вебсайтом, API и других ограничений). Сбор открытых данных осуществляется автоматической программой (парсером), которая обращается к веб-серверу для запроса данных и их последующей обработки.

Открытые данные представляют собой информацию в машиночитаемом формате, которая может свободно и бесплатно использоваться, перерабатываться и распространяться. Данные, полученные путем авторизации на веб-

ресурсах, защищенные авторскими правами и не подлежащие свободному распространению, открытыми не являются. Если технически и существует способ извлечения, возможность использования в исследованиях должна согласовываться с источником.

Веб-скрейпинг обладает комплексом преимуществ и ограничений, сгруппированных вокруг методологических, технических, правовых, финансовых факторов.

1. Возможности веб-скрейпинга.

1.1. Методологические.

Полнота данных. С использованием веб-скрейпинга у социолога появляется возможность однозначно выделить генеральную совокупность объекта исследования при условии, что интернет-данные достаточны для проработки исследовательской проблемы. Становится выполнимым выделение списка всех контрагентов, их индивидуальных характеристик, оценка интернет-рынка – другими методами получить подобные данные без прямого взаимодействия с собственниками платформ невозможно. Открытые данные даже в качестве неосновного источника дополнят описание объекта.

По сравнению с конвенциональными методами сбора данных (опросы, интервью), веб-скрейпинг элиминирует проблему низкого количества откликов в опросах [5. P. 2; 12. С. 163], телефонных звонках и при личных обращениях [13. P. 5].

Качество данных. Скрейпинг, в отличие от канонических способов сбора данных, не связан с отклонениями человеческого фактора: неправильной интерпретацией вопросов интервьюером и респондентом [14. С. 602]. Исключается необходимость валидации интервью на соблюдение программы исследования (корректности отбора респондента и поставленных вопросов, длительности интервью, соответствия ответов вопросам по понятийным критериям) [15. С. 123], допустимого уровня когнитивной нагрузки интервьюера [16. С. 627]. У социолога появляется шанс зафиксировать реальное поведение индивидов, минуя призму восприятия.

Стратегия исследования. Социолог исходит не из теорий, а из доступных данных веб-платформ, которые могут являться маркерами наиболее значимых социальных явлений.

Открывается возможность использовать интернет-данные если не как самостоятельный метод исследования, то как отправную точку для углубления методологии конвенциональных методов сбора данных. Так, анализ распределений генеральной совокупности способствует формированию более надежных стратификационных критериев подбора информантов либо точечной рассылки опросников.

1.2. Технические.

Используя парсер, исследователь налаживает прямую связь с источником данных, создает информационную базу; ее можно обновлять на регулярной основе, отслеживать историю изменений.

1.2. Правовые, этические.

Часто существует возможность получить открытые данные, не нарушая правовые и этические аспекты распространения и использования информации, поскольку четкого регулирования на российском законодательном уровне скрейпинг не имеет. Согласно ст. 5 закона «Об информации, инфор-

мационных технологиях и о защите информации» информация может свободно передаваться в случае отсутствия ограничений к доступу и распространению в других федеральных законах [17. Ст. 5]. В соответствии с этим скрейпинг является законным при соблюдении установленных законодательством РФ ограничений: не нарушаются авторские права; не допускается сбор сведений, составляющих коммерческую тайну; не допускается ограничение конкуренции. Сбор персональных данных пользователей социальных сетей может осуществляться только при согласии [18. Ст. 9]; при этом персональными данными является любая информация, относящаяся прямо или косвенно к субъекту [Там же. Ст. 3].

В международном законодательстве скрейпинг также не получил четкого определения, его юридические рамки прецедентно обрисовываются в судебной практике США (табл. 2).

Таблица 2. **Наиболее заметные судебные дела, связанные с правомерностью скрейпинга**

Длительность судебных тяжб	Истец	Ответчик, осуществляющий скрейпинг	Кого поддержали по решению суда	Решение суда
2000–2009	eBay	Bidder's Edge	Сначала истца, потом ответчика	Высокая активность роботов-парсеров создаёт дополнительную нагрузку к работоспособности сайта eBay; позднее установили, что подобные границы не распространяются к компьютерной среде, поскольку явного ущерба eBay причинено не было
2009	Facebook	Power.com	Истца	Извлечение информации о пользователях – прямое и косвенное нарушение авторских прав
2011–2014	AT&T	Andrew Auernheimer	Истца	Извлечение конфиденциальной пользовательской информации (пусть и общедоступной) несанкционированно
2013	Associated Press	Meltwater	Истца	Нарушается авторское право
2014	QVC	Resutly	Ответчика	Боты ответчика не намеревались нанести ущерб
2017–2020	LinkedIn	HiQ	Ответчика	Скрейпинг правомерен для сбора любых общедоступных данных, не защищённых авторским правом

1.4. Финансово-экономические.

Стоимость сбора данных, по сравнению с конвенциональными методами, может быть ниже, поскольку исключаются затраты на поиск и подбор интервьюеров, оценку их деятельности, поиск информантов, проведение интервью / опросов, аппроксимацию выборочных оценок на генеральную совокупность.

2. Ограничения веб-скрейпинга.

2.1. Методологические.

Объект исследования. Ограниченная область применения: подходит для исследования объектов / явлений, которые могут быть описаны исключительно интернет-данными (блогосфера, взаимодействие через соцсети, цифровые биржи), либо исследование предполагает методику соотнесения онлайн- и офлайн-статусов индивидов [6. С. 30].

Стратегия исследования. Социолог вынужден руководствоваться data-driven подходом: отталкиваться не от теории и операционализации понятий, гипотез, а от данных. Маррес рекомендует на каждом из этапов работы с цифровыми данными обращаться к теоретической рамке и, методично рефлексируя, сравнивать, не отклонился ли исследователь в увлечении методами от своих задач.

Если в конвенциональных методах сбора данных имеется свод правил и предписаний (по расчету выборки, шкалированию), то, применяя скрейпинг, социолог каждый раз должен изобретать новый дизайн исследования согласно структуре данных и контексту платформы (что пользователи пишут на страницах профилей, как общаются) [19. С. 35].

Качество данных. Нечасто имеется возможность выявить методологию заполнения и расчета показателей, интерпретировать пустые значения. Обращение к владельцам платформы при этом не всегда действенно: разработчики могут быть не готовы разглашать алгоритмы, являющиеся интеллектуальной собственностью.

Кроме того, платформы чаще открывают доступ к набору стандартизированных данных, в то время как большая часть неструктурированной информации (коммуникация между пользователями, история посещений) остается закрытой.

Так же как и в конвенциональных социологических методах сбора данных, при скрейпинге сложно однозначно верифицировать достоверность интернет-данных, особенно в условиях тренда на фейк-ньюс, «накрутки» комментариев и голосов, бот-активности. К счастью, часто веб-платформы самостоятельно заботятся о качестве данных, пытаются выявить и предупредить подозрительную активность. Однако в условиях анонимности интернет-среды отсутствуют гарантии того, что одному индивиду соответствует один аккаунт.

Данные многих интернет-платформ содержат как большое количество полезной информации, так и информационного шума (несвязного набора слов, спама, символов выражения эмоциональной окрашенности) – могут потребоваться механизмы фильтрации для его устранения.

2.2. Технические.

Социологии зависимы от проводимых платформой действий: платформа может установить ограничения на пагинацию, объем видимых данных, скорость загрузки во время веб-скрейпинга. Также платформа может внести изменения в структуру данных, что приведет к необходимости преобразования кода либо конфигурации ПО, использующегося для скрейпинга. Исследователь зависит от устойчивости, работоспособности, загруженности платформы.

2.3. Правовые, этические.

Законодательно могут быть предусмотрены ограничения на веб-скрейпинг; его использование может нарушать соглашение об использовании интернет-платформы.

Защищаясь от скрейпинга, платформа может предпринимать действия, ограничивающие возможность извлечения информации, являющейся ее интеллектуальной собственностью. Возникает этический вопрос использования закрытых данных, даже в агрегированном виде.

2.4. Финансово-экономические.

В случаях, когда социолог не обладает компетенциями в программировании, привлечение программиста-подрядчика для скрейпинга может оказаться дороже по сравнению с конвенциональными методами сбора данных.

2.5. Квалификационные.

Существует высокий порог входа в исследование, предполагающее веб-скрейпинг: требуется знание html-разметки, структуры веб-страниц, понимание технических основ сетей, работы интернета, базовое владение языками программирования и работы с базами данных.

Платформы продуцируют множество показателей больших объемов, которые по своей сути являются неструктурированными большими данными: их сбор и анализ требует повышенной когнитивной нагрузки. Результаты энцефалограмм, измерений пульса, индекса утомления и анкетных опросов инженеров-программистов и операторов ПО показывают, что деятельность, связанная с большим количеством функциональных звеньев в системе переработки информации, сопряжена с психической напряженностью и, при отсутствии должного отдыха, прогрессирующим снижением работоспособности [20. С. 27; 21. С. 109].

Научная осмысленность веб-скрейпинга в социологических исследованиях с позиции философии науки

Поскольку веб-скрейпинг, помимо возможностей, предполагает методологические ограничения, связанные с обратной операционализацией понятий, возникает вопрос, насколько теоретически оправданно его использовать. На текущий момент отсутствуют работы, которые давали бы теоретические доказательства научной обоснованности применения веб-скрейпинга в социологических исследованиях с позиции философии науки.

Аргументируем, почему веб-скрейпинг можно считать научно обоснованным инструментом исследовательской оптики социолога.

Веб-скрейпинг – метод цифровой социологии, относящейся к частной теории социологических наук. По теории научных революций [22. P. 7] классическая социология выступает «нормальной наукой», в чью парадигму входят стандартные качественные и количественные методы анализа. Однако в связи с постиндустриальным переходом назревает смена парадигм: онлайн-опросы, средства бизнес-аналитики, потоковые данные, онлайн-платформы и существующие вокруг них научные сообщества способствовали зарождению новой нормальной науки – цифровой социологии.

По Куну, в периоды научных революций разворачивается конкурентная борьба научных сообществ, объединенных разными парадигмами, где победа присуждается сообществу с более устойчивым социально-психологическим настроем и способностью «решать головоломки». Успех научной революции определяется беспрецедентностью теории: может ли привлечь на длительный срок сторонников конкурирующих направлений. Сторонники цифровой социологии только начинают «призывать» на свою сторону технических специалистов (бизнес-аналитиков, UX-исследователей, программистов) для совместного изучения данных цифровых платформ, поэтому говорить о научной революции, заданной вектором цифровой социологии, рано [7. С. 6]. Однако достаточно посмотреть глобальную статистику поисковых запросов к поня-

тиям цифровых данных за последние 10–20 лет – интерес к большим социологическим данным не прекращается: согласно Google Trends всплеск общественного интереса стремится к максимуму в 2020 г. (рис. 1), а в Web of Science устойчиво растет количество цитирований понятия «интернет-данные» в социологии (рис. 2).

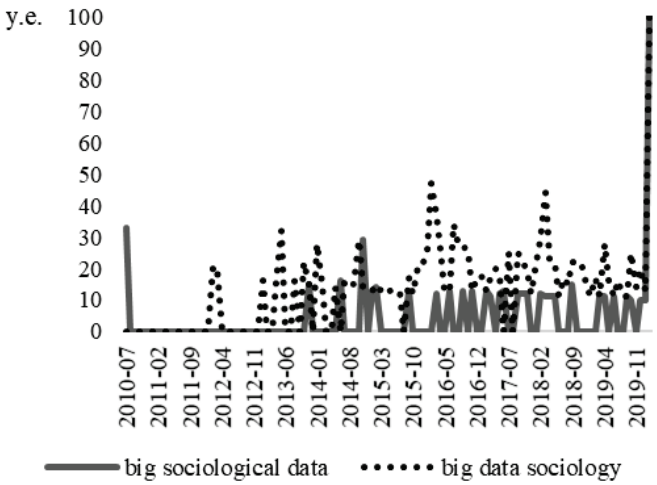

Рис. 1. Уровень интереса к понятию больших социологических данных, у.е. По данным Google Trends на основе статистики поисковых запросов по всему миру; 100 у.е. – максимальный всплеск интереса

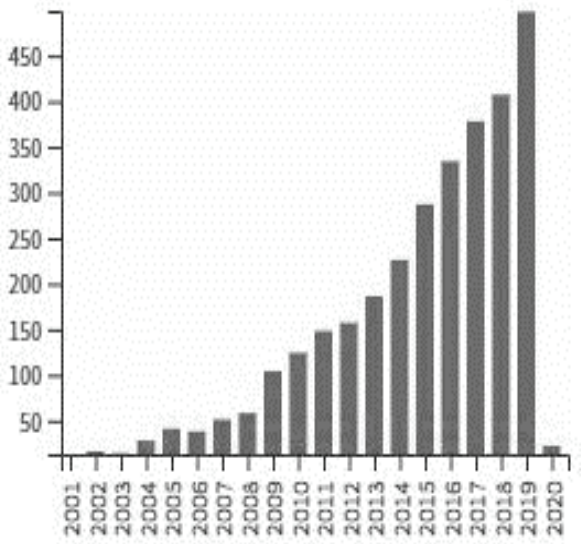

Рис. 2. Число цитирований по теме web data sociology в системе научного цитирования Web of Science, шт.

Дискуссии вокруг веб-скрейпинга справедливы и предсказуемы как для метода нового научного направления и совпадают со всплесками растущего общественного интереса (рис. 3).

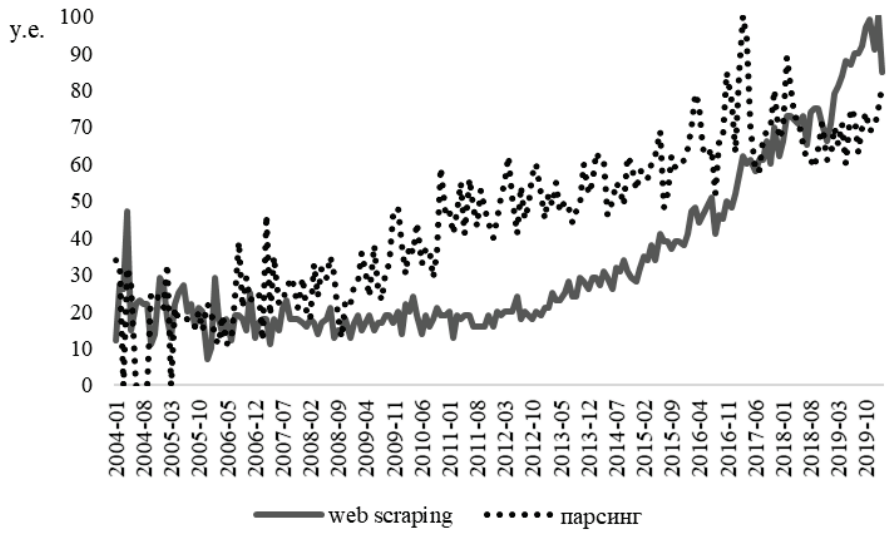

Рис. 3. Уровень интереса к понятиям веб-скрейпинга и парсинга, у.е. По данным Google Trends на основе статистики поисковых запросов по всему миру; 100 у.е. – максимальный всплеск интереса

Веб-скрейпинг – метод, чья ценность в ускорении научной верификации за счет быстрой агрегации данных. Согласно теории логического позитивизма, получившей развитие в рамках Венского кружка, для доказательства научной осмысленности теорий используется верификация – процедура проверки истинности знаний, в ходе которой сложные предложения разделяются на протокольные путем редукции [23. P. 125].

Веб-скрейпинг используется для сбора больших данных, которые детализируются до уровня пользователя / аккаунта. Проверять данные генеральной совокупности, сгруппированные в массиве, на соответствие критериям быстрее и проще, чем осуществлять выборочные проверки аккаунтов. При этом проверка может устанавливаться еще в начале отбора данных (например, когда исследователю требуется отделить только активные аккаунты, аккаунты, созданные в этом году и т.п.). При скрейпинге исследователь осуществляет верификацию протокольных предложений для максимально быстрого обобщения и сведения к теории.

Извлекая данные интернет-платформ, веб-скрейпинг способствует целостному пониманию современной картины мира. Наука предназначена не только для сбора фактического материала, но и формирования целостного мировоззрения [24. С. 19]. Веб-скрейпинг полезен в случаях, когда официальная статистика не помогает описать суть социальных явлений либо отсутствует. Примером группы, для изучения которой веб-скрейпинг представляется единственным источником, являются фрилансеры, осуществляющие поиск контрактов через интернет-биржи труда: только начинают внедряться законодательные инициативы по учету таких работников.

В основе той же синергетики для ученого реальность является миром структур, упорядоченных в соответствии со строгими закономерностями. Интернет-платформы предполагают упорядоченную структуру представления данных (в виде HTML-разметки / прямых запросов через API), разделен-

ных для на простые для восприятия смысловые блоки. Во многом подобная структура и состав информации, представленной на платформе, продиктованы рыночными правилами, а значит, высвечивает наиболее значимые факты.

Веб-скрейпинг – инструмент социолога, готового к междисциплинарному подходу изучения социальной действительности. Порой даже малейшие изменения внешних условий могут приводить к внезапным и радикальным изменениям в системе, которые могут оказать влияние не только на науку, но и на систему подготовки научных кадров. Социолог должен чувствительно реагировать на изменения подходов и методологий, обладать мультидисциплинарными навыками, в том числе использовать программирование для получения доступа к открытой информации.

Однако исследование не должно превращаться в механический анализ данных пользователей: «жестким ядром» [25. С. 1] исследовательской программы социолога должны оставаться стандартные элементы исследования (стратегия, теоретическая рамка, объект, предмет), а «защитным поясом» могут быть цифровая социология, STS и веб-скрейпинг.

Обзор социологических исследований с применением веб-скрейпинга на примере фрилансеров как объекта исследования

Рассмотрим возможности использования интернет-данных, полученных с помощью веб-скрейпинга, на примерах исследований фрилансеров.

По данным крупнейшей российской биржи фрилансеров, собранным веб-скрейпингом, был изучен механизм конкурсов – открытых состязаний между фрилансерами по выполнению задания от заказчика, где за победу присуждается денежное вознаграждение [19. С. 25]. Выборка составила 6 тыс. конкурсов, в которых приняли участие более 300 тыс. фрилансеров. Исследование показывает, что в выборе победителя важное значение отводится репутации фрилансера (рейтингу, отзывам заказчиков на странице профиля) и активному взаимодействию с заказчиком в комментариях. Социально-демографические признаки при этом не являются значимыми и дискриминирующими выбор победителя.

Продолжая тему конкурсов, по данным 30 тыс. проектов биржи freelancer.com были выделены факторы, определяющие победу фрилансера на аукционе с понижением [26. P. 133]. Так, в категории дорогостоящих задач вероятность победы тех назначающих более высокую цену с большим сроком выполнения решения выше и действует в обратном направлении для дешевых задач. Чем сильнее конкуренция за проект, тем ниже конечная стоимость проекта.

Успех будущих проектов фрилансера определяется результатами его предыдущих работ. Так, по данным более миллиона задач, выполненных фрилансерами биржи oDesk, исследователи построили модель предсказания успешности фрилансера по отзывам о его работе на предыдущих проектах [27. P. 1]. Однако не все фрилансеры с равной вероятностью могут стать успешными: на основе анализа 13,5 тыс. профилей фрилансеров бирж TaskRabbit и Fiverr (демографических характеристик, рейтингов и оценок заказчиков, мест фрилансеров в поисковой выдаче) была выявлена расовая и гендерная дискриминация [28. P. 1]. Белые женщины-фрилансеры получают

на 10% оценок меньше, чем мужчины с соответствующим опытом работы, а афроамериканцы в целом обладают более низким рейтингом. Только фрилансеры без фото в профиле ниже по рейтингу, чем афроамериканцы. Отзывы о работе афроамериканских женщин-фрилансеров содержат меньше положительных прилагательных. Среди всех категорий работников мужчины-азиаты располагают наибольшим рейтингом.

Заключение

Веб-скрейпинг обладает потенциалом обогащения социологического знания и вписывается в теоретическую рамку социологических наук. Несмотря на технические и методологические ограничения, стигматизацию интернет-данных, факт появления дискуссий об (анти)научности современных методов цифровой социологии – явление положительное для научного сообщества. Рост знаний достигается в процессе рациональной дискуссии, неизбежно выступающей критикой существующего знания.

Критика конвенциональных источников социологического знания должна привести к популяризации веб-скрейпинга как метода получения информации об объективной действительности, а прогрессивная социология – стать проповедником мультидисциплинарности.

Поддержание подобных тенденций нуждается в содействии научного сообщества. Освоение веб-скрейпинга требует переподготовки социологов на уровне понимания функционирования интернета, усовершенствования навыков программирования и работы с неструктурированной информацией.

Литература

1. *Lupton D.* Digital Sociology: Beyond the Digital to the Sociological // The Australian Sociological Association (TASA) Conference 2013. Melbourne, 2013.

2. *Кравченко С.А.* Новации в социологическом знании: по итогам XIII конференции ECA // Социологические исследования. 2018. № 2. С. 18–24.

3. *Farrell D., Petersen J.* The growth of internet research methods and the reluctant sociologist // Sociological Inquiry. 2010. № 1. P. 114–125.

4. *Petersen J., Farrell D*. Internet Surveys // The Blackwell Encyclopedia of Sociology. New York : John Wiley & Sons, 2016.

5. *Petersen J., Farrell D.* Online Research Methods // The Blackwell Encyclopedia of Sociology. New York : John Wiley & Sons, 2016.

6. *Гришаева С.А., Куликова О.А.* Социально-психологические особенности процесса трансформации социальной структуры общества и процесса коммуникации в цифровом пространстве // Цифровая социология. 2018. № 1. С. 29–34.

7. *Крыштановская О.В.* Бесконтактная социология: новые формы исследований в цифровую эпоху // Цифровая социология. 2018. № 1. С. 4–9.

8. *Marres N.* Digital Sociology: The Reinvention of Social Research. Cambridge : Polity Press, 2017.

9. *Давыдов А.А.* Компьютерные технологии для социологии (обзор зарубежного опыта) // Социологические исследования. 2005. № 1. С. 131–138.

10. *Толстова Ю.Н.* Социология и компьютерные технологии // Социологические исследования. 2015. № 8. С. 3–13.

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

12. *Назарова И.Б.* Непроведение опроса и отказ от интервью // Социологический журнал. 1998. № 1–2. С. 161–167.

13. *Dillman D.A. et al.* Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet // Social science research. 2009. № 1. P. 1–18.

14. *Markou E., Bourgeat E.* Observing the work of interviewers: how the quality of the data collection is constructed // 13th Conference of the European Sociological Association (Un)Making Europe: Capitalism, Solidarities, Subjectivities / ed. by F. Welz. Athens: European Sociological Association, 2017. August 29 – September 1. P. 602.

15. *Берестнева О.Г., Романчуков С.В., Шухарев С.О.* Технология оценки качества работы интервьюеров // Здоровье и образование в XXI веке. 2016. № 3. С. 123–125.

16. *Девятко И.Ф.* Разработка подхода к количественной мультимодальной оценке когнитивной нагрузки интервьюеров: результаты пилотного квазиэксперимента // Вестник РУДН. Социология. 2018. № 4. С. 626–637.

17. *Об информации*, информационных технологиях и о защите информации: федер. закон от 27.07.2006 г. № 149-ФЗ.

18. *О персональных* данных: федер. закон от 27.07.2006 г. № 152-ФЗ.

19. *Стребков Д., Шевчук А., Лукина А., Мелианова Е., Тюлюпо А.* Социальные факторы выбора контрагентов на бирже удаленной работы: исследование конкурсов с помощью «больших данных» // Экономическая социология. 2019. № 3. С. 25–65.

20. *Петрукович В.М., Иванов А.О., Зотов М.В., Федоров С.И.* Влияние гипоксии на умственную работоспособность операторов с различными стратегиями переработки информации в оперативной памяти // Вестник СПбГУ. Социология. 2015. № 3. С. 27–37.

21. *Дружилов С.А.* Психическая напряженность в профессиональной деятельности операторов прокатных станов // Международный журнал прикладных и фундаментальных исследований. Психологические науки. 2014. № 5. С. 109–112.

22. *Kuhn T.* The Structure of Scientific Revolutions. Chicago : University of Chicago Press, 1962.

23. *Carnap R.* Scheinprobleme in der Philosophie. Hamburg : Felix Meiner Verlag, 1929.

24. *Хакен Г.* Тайны природы. Синергетика: учение о взаимодействии. Москва ; Ижевск : Институт компьютерных исследований, 2003.

25. *Лакатос И.* История науки и ее рациональные реконструкции // Из Бостонских исследований по философии науки. М. : Прогресс, 1978. С. 203–235.

26. *Öğüt H.* Factors Affecting Professionals' Selection in High and Low-Value Online Service Procurements // The Service Industries Journal. 2013. № 1 (33). P. 133–149.

27. *Kokkodis M., Ipeirotis P.* Reputation Transferability in Online Labor Market // Management Science. 2016. № 62 (6). P. 1687–1706.

28. *Hannák A. et al.* 2017. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr // ACM Conference on Computer Supported Cooperative Work and Social Computing. 2017.

Olga V. Vilkova, Higher School of Economics (Moscow, Russian Federation).

E-mail: olg.vilkova@gmail.com

Vestnik Tomskogo gosudarstvennogo universiteta. Filosofiya. Sotsiologiya. Politologiya – Tomsk State University Journal of Philosophy, Sociology and Political Science. 2020. 54. pp. 163–175.

DOI: 10.17223/1998863X/54/16

WEB SCRAPING AS A METHOD OF DATA EXTRACTION IN SOCIOLOGICAL STUDIES: ON SCIENTIFIC APPLICABILITY

Keywords: web scraping; digital sociology; sociology of Internet; methods of sociological studies.

The article is devoted to a modern method of data extraction from the Web, web scraping, and its scientific significance and applicability in sociological studies. Based on trends across empirical sociological studies, concepts of digital sociology, science and technology studies (STS), computational sociology, and issues raised at recent international committees' meetings, the current research gives a definition of web scraping and presents an overview of its both methodological and technical opportunities, challenges, and limitations. Advantages and shortcomings are classified across a set of methodological, technical, judicial, ethic, financial, and professional issues and can serve as a perfect framework to be referenced to while weighing risks and rewards at the stage of research design. Comparison with conventional sociological methods, such as survey, in-depth interview or focus group, which lack response rates and have semantic distortions, holds prospects for web scraping as for a method that enables information extraction towards entire population in a timely and structured manner. According to sociology and the philosophy of science, the research aims to determine a place for the method of web scraping in the structure of sociological and scientific knowledge. By alleging to theories of scientific revolutions, science of synergies and the Vienna Circle ideas, the present study tries to prove that,

under the circumstances of a shifting reality, scientific knowledge transforms correspondingly, and research questions imposed to the relevance and scientific meaningfulness of the new theory and its new methods are extremely prompt and expose the necessity of methodology conceptualization. This research is designed to overcome the stigmatization around studies, in which informational bases are mainly constituted by web-platform data. Dealing with online platforms, web scraping is successfully embedded into digital sociology and has potential in covering topics on platform economy. This article urges modern sociologists not to be frightened of learning new instruments and turning their research into interdisciplinary sociological studies. Web scraping benefits interdisciplinary research at the expense of its ability in the simplification of scientific verification processes.

References

1. Lupton, D. (2013) Digital Sociology: Beyond the Digital to the Sociological. *The Australian Sociological Association (TASA) Conference 2013*. Melbourne: [s.n.].

2. Kravchenko, S.A. (2018) Innovations in sociological knowledge: a summary of 13th Conference of ESA. *Sotsiologicheskie issledovaniya – Sociological Studies*. 2. pp. 18–24. (In Russian).

3. Farrell, D. & Petersen, J. (2010) The growth of internet research methods and the reluctant sociologist. *Sociological Inquiry*. 1. pp. 114–125. DOI: 10.1111/j.1475-682X.2009.00318.x

4. Petersen, J. & Farrell, D. (2016a) Internet Surveys. In: Ritzer, G. et al. (eds) *The Blackwell Encyclopedia of Sociology*. New York: John Wiley & Sons.

5. Petersen, J. & Farrell, D. (2016b) Online Research Methods. In: Ritzer, G. et al. (eds) *The Blackwell Encyclopedia of Sociology*. New York: John Wiley & Sons.

6. Grishaeva, S.A. & Kulikova, O.A. (2018) Socio-psychological features of the process of transformation of the social structure of society and the process of communication in the digital space. *Tsifrovaya sotsiologiya – Digital Sociology*. 1. pp. 29–34. (In Russian). DOI: 10.26425/2658-347X-2018-1-29-34

7. Kryshtanovskaya, O.V. (2018) Contactless sociology: new forms of research in a digital age. *Tsifrovaya sotsiologiya – Digital Sociology*. 1. pp. 4–9. (In Russian). DOI: 10.26425/2658-347X-2018-1-4-8

8. Marres, N. (2017) *Digital Sociology: The Reinvention of Social Research*. Cambridge: Polity Press.

9. Davydov, A.A. (2005) Komp'yuternye tekhnologii dlya sotsiologii (obzor zarubezhnogo opyta) [Computer technologies for sociology (a review of foreign experience)]. *Sotsiologicheskie issledovaniya – Sociological Studies*. 1. pp. 131–138.

10. Tolstova, Yu.N. (2015) Sociology and computer technologies. *Sotsiologicheskie issledovaniya – Sociological Studies*. 8. pp. 3–13. (In Russian).

11. Guba, K.S. (2018) Big Data in Sociology: New Data, New Sociology? *Sotsio-logicheskoe obozrenie – Sociological Review*. 1. pp. 213–236. (In Russian). DOI: 10.17323/1728-192X-2018-1-213-236.

12. Nazarova, I.B. (1998) Neprovedenie oprosa i otkaz ot interv'yu [Failure to conduct a survey and refusal of an interview]. *Sotsiologicheskiy zhurnal – Sociological Journal*. 1–2. pp. 161–167.

13. Dillman, D. A. et al. (2009) Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. *Social Science Research*. 1. pp. 1–18. DOI: 10.1016/j.ssresearch.2008.03.007

14. Markou, E. & Bourgeat, E. (2017) Observing the work of interviewers: how the quality of the data collection is constructed. In: Welz, F. (ed.) *13th Conference of the European Sociological Association (Un)Making Europe: Capitalism, Solidarities, Subjectivities*. Athens: European Sociological Association. pp. 602.

15. Berestneva, O.G., Romanchukov, S.V. & Shukharev, S.O. (2016) Tekhnologiya otsenki kachestva raboty interv'yuerov [Technology for assessing the quality of interviewers' work]. *Zdorov'e i obrazovanie v XXI veke – Health and education in the XXI century*. 3. pp. 123–125.

16. Devyatko, I.F. (2018) Developing an approach to multimodal quantitative assessment of interviewers' cognitive load: first results of a field quasi experiment. *Vestnik RUDN. Sotsiologiya – RUDN Journal of Sociology*. 4. pp. 626–637. (In Russian). DOI: 10.22363/2313-2272-2018-18-4-627-637

17. The Russian Federation. (2006a) *Ob informatsii, informatsionnykh tekhnologiyakh i o zashchite informatsii: feder. zakon ot 27.07.2006 g. № 149-FZ* [On information, information technology and information protection: Federal Law No. 149-FZ of July 27, 2006].

18. The Russian Federation. (2006b) *O personal'nykh dannykh: feder. zakon ot 27.07.2006 g. № 152-FZ* [About personal data: Federal Law No. 152-FZ of July 27, 2006].

19. Strebkov, D., Shevchuk, A., Lukina, A., Melianova, E. & Tyulyupo, A. (2019) Social Factors of Contractor Selection on Freelance Online Marketplace: Study of Contests Using "Big Data". *Ekonomicheskaya sotsiologiya – Economic Sociology*. 3. pp. 25–65. (In Russian).

20. Petrukovich, V.M., Ivanov, A.O., Zotov, M.V. & Fedorov, S.I. (2015) Hypoxia influence on the mental working capacity of operators who used different strategies of information processing in a working memory system. *Vestnik SPbGU. Sotsiologiya – Vestnik of Saint-Petersburg University. Sociology*. 3. pp. 27–37. (In Russian).

21. Druzhilov, S.A. (2014) Psikhicheskaya napryazhennost' v professional'noy deyatel'nosti operatorov prokatnykh stanov [Mental tension in the professional activities of rolling mill operators]. *Mezhdunarodnyy zhurnal prikladnykh i fundamental'nykh issledo-vaniy. Psikhologicheskie nauki*. 5. pp. 109–112.

22. Kuhn, T. (1962) *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 1962.

23. Carnap, R. (1929) *Scheinprobleme in der Philosophie*. Hamburg: Felix Meiner Verlag.

24. Haken, G. (2003) *Tayny prirody. Sinergetika: uchenie o vzaimodeystvii* [Secrets of Nature. Synergetics: The Doctrine of Interaction]. Translated from English. Moscow; Izhevsk: Institut komp'yuternykh issledovaniy.

25. Lakatos, I. (1978) Istoriya nauki i ee ratsional'nye rekonstruktsii [History of science and its rational reconstruction]. In: Gryaznov, B.S. & Sadovsky, V.N. (eds) *Struktura i razvitie nauki. Iz Bostonskikh issledovaniy po filosofii nauki* [Structure and Development of Science. From Boston Studies on the Philosophy of Science]. Moscow: Progress. pp. 203–235.

26. Öğüt, H. (2913) Factors Affecting Professionals' Selection in High and Low-Value Online Service Procurements. *The Service Industries Journal*. 1(33). pp. 133–149. DOI: 10.1080/02642069.2011.600445

27. Kokkodis, M. & Ipeirotis, P. (2016) Reputation Transferability in Online Labor Market. *Management Science*. 62(6). pp. 1687–1706. DOI: 10.1287/mnsc.2015.2217

28. Hannák, A. et al. (2017) Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr. *ACM Conference on Computer Supported Cooperative Work and Social Computing*.