

## НАУЧНАЯ СТАТЬЯ

DOI: 10.21684/2587-8484-2019-3-3-74-84

УДК 316.325

## Big Data как источник социологической информации: пример анализа блога губернатора Петербурга

Анна Васильевна Мальцева<sup>1</sup>, Михаил Сергеевич Матвеев<sup>2</sup>,  
Мария Борисовна Моисеева<sup>3</sup>

<sup>1</sup> доктор социологических наук, доцент кафедры социального анализа и математических методов в социологии, Санкт-Петербургский государственный университет (г. Санкт-Петербург, РФ)  
ORCID: 0000-0003-1322-6255; Web of Science ResearcherID: K-2094-2013  
a.maltseva@spbu.ru

<sup>2</sup> студент Санкт-Петербургского государственного университета (г. Санкт-Петербург, РФ)  
ORCID: 0000-0002-5378-6559; Web of Science ResearcherID: C-8797-2019  
mikhail.matveev97@gmail.com

<sup>3</sup> студентка Санкт-Петербургского государственного университета (г. Санкт-Петербург, РФ)  
mariamoiseeva108@gmail.com

**Аннотация.** В статье исследуются возможности Big Data (больших данных) в построении прогнозов, социальные сети как источник информации об обществе, а также рассматривается роль технологии больших данных в процессах обеспечения безопасности. Целью исследования является изучение возможности получения социально значимой информации посредством больших данных, в частности, по одному из аспектов экологической безопасности. Рассмотрены фундаментальные социологические теории, такие как теории общества риска У. Бека, Э. Гидденса, а также информационного общества, учитывающие особенности влияния новейших информационных технологий на общество, в том числе больших данных. На основе ряда исследований, систематизирующих подходы к определению больших данных, выводится авторское определение данного понятия. В работе использованы качественные методы, в частности контент-анализ. Анализ был произведен на примере предоставленного Комитетом по информатизации и связи города Санкт-Петербурга приложения «Инцидент-менеджмент». Эмпирическая база исследования — 16 694 комментария на публичной странице губернатора Санкт-Петербурга Александра Беглова, полученные посредством программы на языке Python с использованием VK Api. В ходе исследования проанализированы комментарии пользователей на тему экологических рисков по районам Санкт-Петербурга, выявлены районы, подверженные экологическим рискам в наибольшей степени, а также источники этих рисков и субъекты ответственности за них. В результате проведенного исследования было сделано заключение, что неформальные средства коммуникации представителей исполнительной власти и граждан вызывают большой интерес в качестве источника социологической информации, исследуемого с помощью алгоритмов и методов анализа больших данных. Такой анализ помогает повысить уровень безопасности за счет снижения неопределенности и получения новых знаний об обществе. С другой стороны, исследователи выявили возможные социальные риски, связанные со сбором, хранением и применением больших данных. В частности, это риски внешнего вмешательства в работу приложения. На практике эти риски должны быть предупреждены и минимизированы путем совершенствования и тестирования приложений компанией-разработчиком. Социология, в свою очередь, должна учитывать этот факт и избегать внесения неправомерных погрешностей в исследовательские результаты.

**Ключевые слова:** большие данные, безопасность, экологическая безопасность, риск, контент-анализ, информационные технологии, информация, социальные медиа.

## ВВЕДЕНИЕ

Актуальность данной работы заключается в том, что в последнее время термин «большие данные», или Big Data, стал популярен в академических и профессиональных кругах и повсеместно используется в средствах массовой информации. При этом, по словам А. де Мауро, термин употребляется во множестве различных контекстов и часто не соответствует настоящему смыслу больших данных, что мешает полно и эффективно их исследовать [12]. Тем не менее, принимая во внимание специфику роста технического прогресса и то, что цифровые технологии проникают в подавляющее большинство аспектов жизни современного человека, использование данных — характерная черта современного общества. Каждый день в мире производится 2,5 триллиона мегабайт данных. Появляется возможность анализировать данные с помощью специфических алгоритмов аналитики или методов и использовать их, например, для прогнозирования, получения информации о рисках и угрозах для обеспечения безопасности, которая способствует устойчивому развитию общества. Однако помимо положительных эффектов любые технологии несут потенциальные риски и угрозы для общественной безопасности, которые рассматривают многие социологи, среди них Э. Гидденс и У. Бек.

С точки зрения социологии, не только явление, но и сам обозначающий его термин может вызывать интерес. По мнению исследователей, он также оказывает влияние на общество: популярный дискурс о больших данных способствует развитию тенденции представлять это явление как имеющее колоссальное значение для любых коммерческих, управленческих, правительственных и исследовательских целей, что уже сейчас вызывает неправильное (perverse) представление о безграничных возмож-

**Цитирование:** Мальцева А. В. Big Data как источник социологической информации: пример анализа блога губернатора Петербурга / А. В. Мальцева, М. С. Матвеев, М. Б. Моисеева // Siberian Socium. 2019. Том 3. № 3 (9). С. 74-84.

DOI: 10.21684/2587-8484-2019-3-3-74-84

ностях больших данных [17]. А социологи, занимающиеся вопросами общественного мнения, заявляют, что присущие большим данным характеристики дают более эффективные возможности для построения прогнозов, чем традиционные методы [7, с. 30-43]. В частности, Д. Бойд и К. Кроуфорд пишут, что большие данные создают радикальный сдвиг в том, как «мы думаем об исследованиях», и глубокие изменения на уровне эпистемологии и этики. Большие данные переопределяют ключевые вопросы о конституции знаний [10, с. 662-679]. При этом исследователи выделяют и существующие проблемы, например репрезентативности, в работе с ними [5, с. 141-145]. То есть большие данные в контексте как науки, так и безопасности позволяют получить новую информацию об обществе.

Таким образом, цель данного исследования заключается в том, чтобы определить потенциал больших данных в получении социологической информации на примере сведений об экологической безопасности. Возможности обеспечения безопасности на основе больших данных были исследованы с помощью приложения «Инцидент-менеджмент» — системы реагирования на публикации и жалобы в социальных сетях, использующейся в государственном управлении. В работе были использованы качественные методы, в частности контент-анализ. Эмпирическая база исследования — комментарии на публичной странице и. о. губернатора Санкт-Петербурга Александра Беглова, полученные посредством программы на языке Python с использованием VK Api. Она включает 16 694 комментария. В ходе исследования также проанализированы комментарии пользователей на тему экологических рисков по районам Санкт-Петербурга.

## ОСНОВНАЯ ЧАСТЬ

Технологии, в том числе большие данные, с социологической точки зрения можно изучать в контексте влияния на общество. Так, М. Кастельс характеризует современное общество как «информациональное». Знания и информация были центральными во всех обществах, однако благодаря развитию технологий на сегодняшний день информация стала иметь определяющее значение в обществе. М. Кастельс уверен, что термин «безопасность» нуждается в переосмыслении в условиях новой социально-экономической реальности [3, с. 319].

Э. Гидденс понимает «национальные государства» как информационные общества, поскольку государственная власть подразумевает рефлексивный сбор, хранение и управление информацией. Они являются высокоорганизованными системами, и поэтому органы государственного управления нуждаются в высоком уровне информационного обеспечения, а для этого необходимы новые технологии [13, с. 178].

Далее нужно остановиться на концепции «общества риска» У. Бека. Это общество, которое систематически сталкивается с «опасностями и ненадежностью, возникающими и создаваемыми самим научно-техническим прогрессом» [1]. Последующие исследователи, в частности К. Керр, говорят о том, что в теории Бека прогнозирование и риск являются взаимосвязанными понятиями: страх перед последним ведет к возрастанию желания обладать способностями к прогнозу и к попыткам блокировать риски до того, как они произойдут. А Э. Гидденс пишет, что современность основана на выявлении и предотвращении риска с помощью сбора информации [2]. Также ученый говорит, что особенностью модернизации за последние несколько десятилетий стало быстрое цифровое и технологическое развитие и рационализация, что теперь ассоциируется с растущей угрозой искусственных рисков, создаваемых самим обществами. Эти риски связаны со значительным уровнем человеческой активности, в отличие от внешних рисков, таких как стихийные бедствия, которые были главной заботой предыдущей индустриальной эры [14, с. 1-10].

Безопасность определяется федеральным законом Российской Федерации от 5 марта 1992 г. № 2464-1 как «состояние защищенности жизненно важных интересов личности, общества и государства от внутренних и внешних угроз» [9]. В новой редакции закона от 2010 г., а также в Стратегии национальной безопасности Российской Федерации до 2020 г. стратегическими являются задачи важнейших социальных, политических и экономических преобразований для создания безопасных условий реализации конституционных прав и свобод граждан Российской Федерации, осуществления устойчивого развития страны, сохранения территориальной целостности и суверенитета государства [8]. В. Кузнецов выделял следующие виды безопасности: духовная, экономическая, военная, экологическая, информационная, политическая, социальная [4, с. 147]. В качестве «средств обеспечения» безопасности могут быть использованы информационные технологии, в частности большие данные.

В результате анализа ряда исследований нами была выведена обобщенная дефиниция, в соответствии с которой под большими данными понимаются объемы данных, важнейшими параметрами которых выступают скорость и исчерпанность. Получение из них информации и знаний требует применения специальных технологий и аналитических методов. В основу легло определение А. де Мауро [12], составленное на основе анализа аннотаций научных работ, скорректированное с учетом правок Р. Китчина [15].

Скорость считается ключевым атрибутом больших данных. Большие данные создаются непрерывно. К примеру, данные могут создаваться в то время, когда пользователь просматривает веб-сайты. Исчерпываемость заключается в том, что большие данные стремятся охватить всю совокупность ( $n = \text{всё}$ ) внутри системы, а не выборку. Например, Twitter захватывает все твиты, сделанные всеми аккаунтами, а не образец твита.

Также А. де Мауро выделил четыре главных аспекта больших данных: информация, технологии, методы, влияние на общество — и сделал вывод о том, что они являются наиболее распространенными темами в существующей академи-

ческой литературе, посвященной большим данным [11, с. 122-135].

Первый аспект — информация. Данные становятся доступными благодаря их оцифровке, что делает их воспринимаемыми для компьютера, помимо этого существует большое разнообразие форматов. Также существует огромное количество устройств, которые соединены Интернетом. Из этого объема данных можно извлечь информацию, согласно иерархии «данные — информация — знание». Именно информация является целью анализа данных. Работа с большими данными подразумевает требования к технологическим вычислительным ресурсам и к хранилищам данных. В частности, технология Hadoop позволяет осуществлять распределенную обработку больших объемов данных с использованием группы рассредоточенных машин. Также важным аспектом является возможность передачи данных и хранения огромных объемов данных.

Для извлечения информации применяется комплекс методов их обработки и анализа. Это могут быть традиционные методы (корреляция, контент-анализ) либо инновационные (обработка естественного языка, нейронные сети и другие).

При этом использование и управление большими данными влияет на многие сферы жизни общества. Благодаря технологиям расширяется диапазон источников, позволяющих получить новые знания, более полную картину социального, в первую очередь благодаря свойству исчерпанности. Положительные черты больших данных вытекают из их главных характеристик: исчерпанность позволяет анализировать весь массив данных, а не выборку, скорость способствует актуальности информации, поскольку большие данные создаются непрерывно. Создание больших данных дает возможность увеличить степень полноты, точности и актуальности информации, что способствует минимизации рисков.

Однако большие данные становятся источниками новых рисков, в том числе возникают следующие проблемы.

Во-первых, появляются проблемы конфиденциальности и защиты частной информации, ограничения осведомленности субъектов данных

о том, что данные о них собираются и изучаются — это становится возможным за счет того, что правительства имеют право собирать информацию без заключения частных соглашений с согласия пользователя. Также к этим проблемам относится симбиоз государственного и частного сектора, потому что многие из функций, которые мы воспринимаем как функции государственного сектора (подсчет, классификация и обслуживание наших потребностей как граждан), фактически выполняются частным сектором, что имеет соответствующие последствия для прозрачности и подотчетности.

Во-вторых, проблемы возникают на разных этапах работы с большими данными. Например, на этапах сбора и анализа данных из-за незнания технологических принципов работы с большими данными, выбора неподходящих методов или неграмотного их использования результаты применения больших данных могут оказаться ошибочными. Также сложности могут возникнуть на этапе интерпретации данных из-за чрезмерного доверия людей к технологиям больших данных и, как следствие, привести к принятию неверных управленческих решений. Ярким примером может служить проект Google Flu Trends (GFT), целью которого было прогнозирование эпидемии гриппа, основываясь на поисковых запросах людей о лекарствах. Однако в феврале 2013 г. проект потерпел неудачу, показав более чем удвоенное количество посещений врачей по поводу гриппа, чем Центры по контролю и профилактике заболеваний (CDC), что вызвало активный дискурс в различных медиа о больших данных [16, с. 1203-1205].

В-третьих, существует проблема необходимости вложения крупных средств в технологический сектор, а также наличия примеров, когда эти инвестиции не привели к ожидаемым результатам.

Примером реального внедрения технологии больших данных в деятельность государственных органов Российской Федерации может служить «Инцидент-менеджмент» — новая система реагирования на публикации в социальных сетях (жалобы, вопросы, обращения). Разработчиком программы является Общество с ограниченной

ответственностью «Медиалогия». Компания осуществляет сбор данных внутри социальных сетей (Facebook, «ВКонтакте», «Одноклассники», Instagram, YouTube, Twitter, Google+, «Ответы Mail.ru», LiveJournal) форумов и СМИ [6]. На серверы компании загружается более 100 миллионов сообщений в сутки посредством инструментов Apache Flume, Apache Kafka, Apache Ignite и других. Поточковая обработка осуществляется с помощью компьютерной лингвистики, кластеризации и машинного обучения. На трех дата-центрах общей вместимостью 650 терабайтов находится так называемое «озеро данных» (data lake). Обработка и анализ данных происходит на основе Apache Big Data Stack. Приложения для клиентов получают данные на основе этой цепочки, которая осуществляется внутри компании. В частности, так работают «Инцидент-менеджмент» и другие решения для анализа социальных сетей и медиа.

Приложение осуществляет автоматизированный мониторинг сообщений и комментариев. После их обработки на основе отобранных создаются инцидент-сообщения, или сообщения, взятые в работу. В дальнейшем инцидент распределяется на ответственных и исполнителей внутри системы. Это могут быть представители органов исполнительной власти или местного самоуправления. На сегодняшний день программа работает в тестовом режиме и осуществляет анализ на основе только публичной страницы исполняющего обязанности губернатора Александра Беглова. Однако в перспективе данное приложение способствует улучшению функционирования общества с помощью правильного и своевременного получения информации о потребностях и проблемах. Также способствует снижению срока реагирования на основе полученных фактов.

Далее нами было проведено эмпирическое исследование, целью которого являлось изучение возможности получения информации посредством технологий больших данных для обеспечения безопасности (на примере экологической безопасности). Экологическая безопасность — это защищенность среды обитания от угроз, соз-

даваемых деятельностью человека. Экологическая безопасность современного мегаполиса — это сложный и многоаспектный процесс. Е. Г. Тихомирова и Е. Г. Сёмин систематизировали мероприятия по минимизации экологических рисков и угроз на ряд блоков. В данной работе исследуются санитарный блок (вопросы уборки и очистки городского пространства и входящих в нее акваторий от мусора и отходов) и деконтаминационный блок (их дальнейшая транспортировка и переработка).

Посредством специально разработанной программы на языке Python с помощью VK API был выгружен массив из 16 694 комментариев под постами Александра Беглова за март 2019 г. Из них было выделено 145 комментариев, в которых упоминались проблемы, связанные с мусором и другими коммунальными отходами. Далее с помощью контент-анализа было изучено, какую информацию можно получить из источника для последующей реакции на возникающие риски, а также проведен анализ мнения жителей по поводу экологических угроз, связанных с мусором. Так как само приложение пока работает в тестовом режиме, то и анализ был осуществлен на основании того же самого источника.

Контент-анализ был осуществлен для выявления информации по трем направлениям: проблемы, связанные с уборкой и переработкой мусора; локализация экологических рисков по районам; оценка жителями города деятельности социальных субъектов (см. таблицу 1).

Далее представлены результаты проведенного эмпирического исследования по каждой категории анализа:

1. *Проблемы, связанные с уборкой и переработкой мусора.* Распределение комментариев пользователей в зависимости от того, какие проблемы были ими указаны, представлены в таблице 2. Благодаря мониторингу социальных сетей, можно получить информацию или уточняющие сведения об одном из видов экологических проблем — проблемах с мусором. Деконтаминационный блок, содержащий 32% от всех комментариев, был поделен на два дополнительных блока. 16% от всех комментариев связаны со сбором,

вывозом и хранением мусора: переполненность мусорных контейнеров, отсутствие баков на площадках, проблемы раздельного сбора мусора. Еще 16% включают жалобы на мусорные полигоны, несанкционированные свалки, мусоросжигательные и мусороперерабатывающие заводы.

Санитарный блок составил 68% от общего массива. 44% от всех комментариев содержат только упоминание о проблеме, без указания характера или других дополнительных свойств, из них 30% содержали словосочетание «горы

мусора». Таким образом, дополнительную информацию из этих данных извлечь не представляется возможным. Также отдельно авторы комментариев выделяют проблемы, связанные с мусорным загрязнением внутри городских акваторий, парков, пляжей, заброшенных зданий, то есть указывают среду загрязнения — оставлено 9% такого рода комментариев. В 15% комментариев даны дополнительные характеристики: сопутствующее пылевое загрязнение, строительный мусор, различные животные и организмы.

Таблица 1. Контент-анализ интернет-источника  
Table 1. Content analysis of the Internet source

Категории анализа	Единицы анализа	Единицы счета
Упоминание проблем, связанных с уборкой и переработкой мусора	Конкретные адреса и названия мест, в которых есть угроза экологической безопасности	Слова и словосочетания
Упоминание районов, где локализованы экологические риски	Слова «мусор» и семантически связанные с ним слова	
Упоминание социальных субъектов и оценка их деятельности	Конкретные имена социальных субъектов, должности и другие характеристики	

Таблица 2. Информация о распределении комментариев пользователей о проблемах, связанных с уборкой и переработкой мусора  
Table 2. Information on the distribution of user comments on problems related to garbage collection and recycling

Деконтаминационный блок (32%)		Санитарный блок (68%)		
Сбор, вывоз, хранение мусора (16%)	Свалки, полигоны, заводы (16%)	Дополнительные характеристики (15%)	Места загрязнения (9%)	Нет уточнения (44%)

Таблица 3. Информация о распределении комментариев пользователей об оценке деятельности социальных субъектов  
Table 3. Information on the distribution of user comments on the evaluation of social actors

Указывают социальных субъектов (33%)	Нейтральные (48%)	А. Беглов (83%)
		Другие (17%)
	Негативные (51%)	Чиновники, представители власти (39%)
		ЖКС (26%)
		Губернаторы СПб (13%)
Положительные (1%)	Другие (26%)	
Не указывают социальных субъектов (67%)		

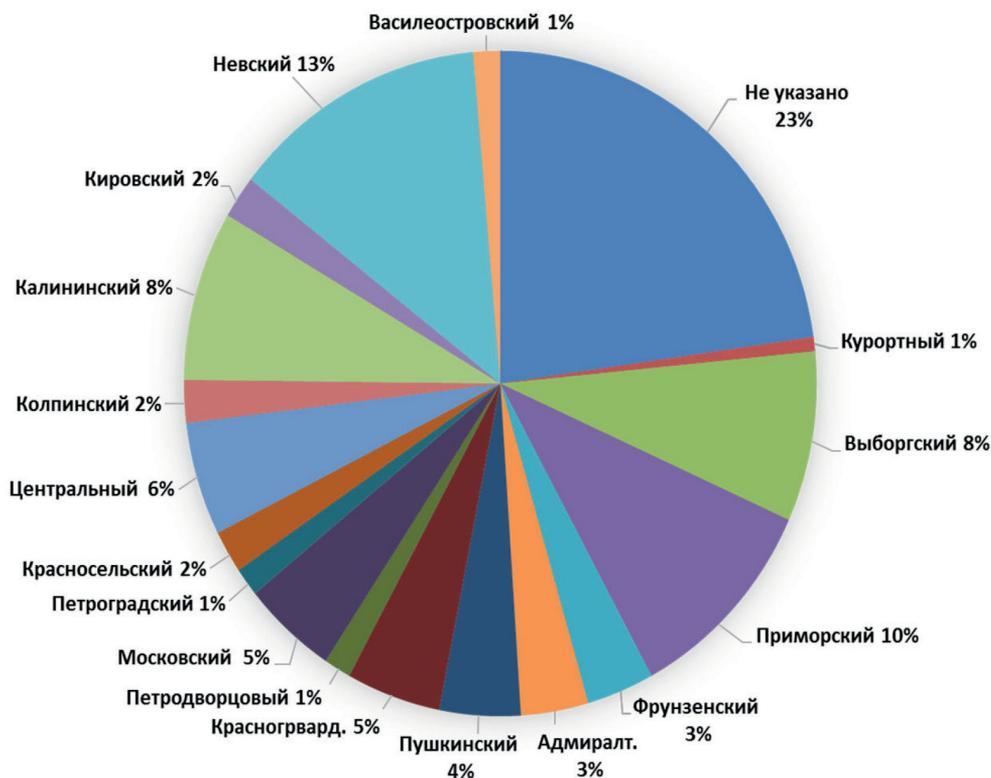
2. *Информация об экологических рисках по районам Санкт-Петербурга.* Исходя из анализа такого источника, можно сделать вывод о том, что из 77% комментариев можно было получить информацию о местоположении источников экологических угроз, связанных с мусором. Более детализированное распределение указанных районов представлено на рис. 1.

3. *Оценка деятельности социальных субъектов.* Помимо этого, в ходе анализа были изучены упоминания о социальных субъектах и оценка их деятельности в контексте обеспечения экологической безопасности (см. таблицу 3). Все упоминания были разделены на три категории: положительные, нейтральные — если в комментариях содержится просьба или обращение без оценочных суждений, отрицательные — критикующие

действия тех акторов, которые, по мнению людей, несут ответственность за проведение комплекса необходимых для улучшения ситуации мер.

Из 145 комментариев было зафиксировано 48 комментариев (33%), указывающих на соц. субъекта. Из них 48% содержат упоминание или просьбу, без позитивных или отрицательных оценок. 80% процентов из нейтральных высказываний относятся к и. о. губернатора А. Беглову. 51% от общего числа содержат негативные упоминания. Наибольшее недовольство вызывает работа городских чиновников и властей — 39% от всего числа отрицательных комментариев, а также специализированных служб (ЖКС) — 26%. Кроме того, в негативном тоне описывается опыт предыдущих глав Санкт-Петербурга — 13%.

Рис. 1. Информация о распределении комментариев пользователей об экологических рисках по районам Санкт-Петербурга  
Fig. 1. Information on the distribution of user comments on environmental risks by districts of St. Petersburg



### ЗАКЛЮЧЕНИЕ

Итак, можно сделать следующие выводы по эмпирической части статьи:

1. Наиболее подверженные экологическим рискам районы Санкт-Петербурга — Невский, Приморский, Выборгский и Калининский.
2. Самые актуальные проблемы связаны с мусором в городском пространстве, а также с его сбором, вывозом, хранением и переработкой.
3. Больше всего недовольства вызывает работа городских чиновников и представителей власти, при этом просьбы и призывы справиться с ситуацией в основном обращены к А. Беглову.
4. Из текстового анализа пользовательских данных можно сделать вывод, что часть комментариев не содержит уточнений о проблеме (44%), указания места загрязнения (23%) или оценки деятельности социальных субъектов (67%).
5. Таким образом, приложение «Инцидент-менеджмент» и на данный момент способствует получению информации для обеспечения безопасности. В дальнейшем выборка из нескольких социальных сетей и использование широкого диапазона методов позволит получать более полные данные.

В результате проведенной работы можно сформулировать общие выводы. Выявлено, что фун-

даментальные социологические теории, такие как теории общества риска и информационного общества, учитывают особенности влияния на общество новейших информационных технологий, в том числе больших данных.

На основе ряда исследований, систематизирующих подходы к определению больших данных, выведено авторское определение данного понятия. Для изучения роли программы «Инцидент-менеджмент», использующей технологии больших данных, в экологической безопасности было применено сочетание традиционного контент-анализа и возможностей, предоставляемых трендовыми языками программирования (Python).

Контент-анализ социального медиа «ВКонтакте», а именно страницы Александра Беглова, продемонстрировал возможности получения новой информации и знаний об угрозах и рисках, которые необходимо минимизировать для стремления к стабильному развитию общества, и подтвердил выдвинутую гипотезу.

Благодаря теориям, названным выше, были учтены возможные риски, которые несли «Инцидент-менеджмент» и ее компания-разработчик, например, возможность утечки информации.

Несмотря на доказанный высокий потенциал данных технологий, необходимо продолжить исследовать риски, которые они могут нести, и проводить работу по их минимизации. Поэтому представляется необходимым дальнейшее изучение подобных приложений.

### СПИСОК ЛИТЕРАТУРЫ

1. Бек У. Общество риска. На пути к другому модерну / У. Бек. М.: Прогресс-Традиция, 2000. 384 с.
2. Гидденс Э. Устроение общества: очерк теории структуризации / Э. Гидденс. 2-е изд. М.: Академический Проект, 2005. 528 с.
3. Кастельс М. Галактика Интернет. Размышления об Интернете, бизнесе и обществе / М. Кастельс. Екатеринбург: У-Фактория, 2004. С. 319.
4. Кузнецов В. Социология безопасности: учебное пособие / В. Кузнецов. М.: МГУ, 2007. 423 с.
5. Мальцева А. В. Проблемы репрезентативности при работе с «большими данными» / А. В. Мальцева // Социальные практики и управление: проблемное поле социологии: материалы Сибирского социологического форума с международным участием / отв. ред. С. В. Ровбель, С. А. Ильиных. Новосибирск: НГУЭУ, 2017. С. 141-145.
6. Медиалогия: инструмент оценки эффективности коммуникаций. Как работает. URL: <https://tech.mlg.ru/>
7. Одинцов А. В. Социология общественного мнения и вызов Big Data / А. В. Одинцов // Мониторинг общественного мнения: экономические и социальные перемены. 2017. № 3. С. 30-43.

8. Стратегия национальной безопасности Российской Федерации до 2020 года (утв. Указом Президента Российской Федерации от 12 мая 2009 года № 537). URL: <http://kremlin.ru/supplement/424>
9. Федеральный закон «О безопасности» от 28.12.2010 № 390-ФЗ // Справочно-правовая система «Консультант Плюс». URL: [http://www.consultant.ru/document/cons\\_doc\\_LAW\\_108546/](http://www.consultant.ru/document/cons_doc_LAW_108546/)
10. Boyd D. Critical questions for BigData / D. Boyd, K. Crawford // *Information, Communication and Society*. 2012. Vol. 15. Issue 5. Pp. 662-679. DOI: 10.1080/1369118X.2012.678878
11. De Mauro A. Formal definition of Big Data based on its essential features / A. De Mauro, M. Greco, M. Grimaldi // *Library Review*. 2016. Vol. 65. No 3. Pp. 122-135. DOI: 10.1108/LR-06-2015-0061
12. De Mauro A. What is big data? A consensual definition and a review of key research topics / A. De Mauro, M. Greco, M. Grimaldi // Conference: 4th International Conference on Integrated Information. Madrid, 2014. URL: [https://www.researchgate.net/publication/265775800\\_What\\_is\\_Big\\_Data\\_A\\_Consensual\\_Definition\\_and\\_a\\_Review\\_of\\_Key\\_Research\\_Topics](https://www.researchgate.net/publication/265775800_What_is_Big_Data_A_Consensual_Definition_and_a_Review_of_Key_Research_Topics) DOI: 10.13140/2.1.2341.5048
13. Giddens A. *The Nation State and Violence: Volume Two of a Contemporary Critique of Historical Materialism* / A. Giddens. University of California Press, 1987. P. 178.
14. Giddens A. Risk and Responsibility / A. Giddens // *The Modern Law Review*. 1999. Vol. 62. No 1. Pp. 1-10. DOI: 10.1111/1468-2230.00188
15. Kitchin R. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets / R. Kitchin, G. McArdle // *Big Data & Society*. 2016. Pp. 1-10. DOI: 10.1177/2053951716631130
16. Lazer D. The parable of Google Flu: traps in Big Data analysis / D. Lazer, R. Kennedy, G. King, A. Vespignani // *Science*. 2014. 14 Mar. Vol. 343. Issue 6176. Pp. 1203-1205. DOI: 10.1126/science.1248506
17. Lupton D. The Thirteen Ps of Big Data / D. Lupton // *This Sociological Life: a blog by sociologist Deborah Lupton*. 2015. URL: <https://simplysociology.wordpress.com/2015/05/11/the-thirteen-ps-of-big-data/>

## RESEARCH ARTICLE

DOI: 10.21684/2587-8484-2019-3-3-74-84

UDC 316.325

## Big Data as a source of sociological information: the analysis of the blog by Saint Petersburg's governor

Anna V. Maltseva<sup>1</sup>, Mikhail S. Matveev<sup>2</sup>, Maria B. Moiseeva<sup>3</sup>

<sup>1</sup> Dr. Sci. (Soc.), Associate Professor,  
Department of Social Analysis and Mathematical Methods on Sociology,  
Saint Petersburg State University (Saint Petersburg, Russian Federation)  
ORCID: 0000-0003-1322-6255; Web of Science ResearcherID: K-2094-2013  
a.maltseva@spbu.ru

<sup>2</sup> Undergraduate Student, Saint Petersburg State University (Saint Petersburg, Russian Federation)  
ORCID: 0000-0002-5378-6559; Web of Science ResearcherID: C-8797-2019  
mikhail.matveev97@gmail.com

<sup>3</sup> Undergraduate Student, Saint Petersburg State University (Saint Petersburg, Russian Federation)  
mariamoiseeva108@gmail.com

**Abstract.** This article explores the possibilities of Big Data in forecasts and social networks as a source of information about society, as well as its role in security processes. The authors aim to study the possibility of obtaining socially significant information through Big Data, in particular, on one of the aspects of environmental safety. Fundamental sociological theories are considered, such as theories of the risk society by W. Beck and E. Giddens, as well as the information society (accounting for the peculiarities of the influence of the latest information technologies on society, including Big Data). Based on a number of studies systematizing approaches to the definition of Big Data, the authors derive their own. Using qualitative methods (in particular, content analysis), the authors analyzed the example of the incident-management application provided by the Committee on Informatization and Communications of the City of Saint Petersburg. The empirical base of the study includes 16,694 comments on the public page of the governor of St. Petersburg Alexander Beglov, obtained through a program in Python and using VK Api. The study analyzes user comments on environmental risks in the districts of St. Petersburg, identifies areas that are most exposed to environmental risks, as well as the sources of these risks and the subjects of responsibility for them. The results show that informal means of communication between representatives of the executive branch and citizens are of great interest as a source of sociological information investigated using algorithms and methods for analyzing Big Data. Such analysis helps in increasing security by reducing uncertainty and gaining new knowledge about society. On the other hand, the researchers have identified possible social risks associated with collecting, storing, and using Big Data. In particular, these are the risks of external interference with the application. In practice, these risks should be prevented and minimized by improving and testing applications by the developing company. Sociology, in turn, must consider this fact and avoid introducing undue errors in the research results.

**Keywords:** Big Data, security, ecological security, risks, content analysis, information technologies, information, social networks.

**Citation:** Maltseva A. V., Matveev M. S., Moiseeva M. B. 2019. "Big Data as a source of sociological information: the analysis of the blog by Saint Petersburg's governor". *Siberian Socium*, vol. 3, no 3 (9), pp. 74-84. DOI: 10.21684/2587-8484-2019-3-3-74-84

## REFERENCES

1. Beck W. 2000. Risk Society. On the Way to Another Art Nouveau. Moscow: Progress-Tradition. [In Russian]
2. Giddens E. 2005. Organization of Society: Essay on the Theory of Structure. 2<sup>nd</sup> edition. Moscow: Academic Project. [In Russian]
3. Castells M. 2004. Galaxy Internet. Reflections on the Internet, Business and Society. U-Factoria. [In Russian]
4. Kuznetsov V. 2007. Sociology of Security: A Training Manual. Moscow: Moscow State University. [In Russian]
5. Maltseva A. V. 2017. "The problems of representativeness when working with 'big data': Social practices and management: the problem field of sociology". Proceedings of the Siberian sociological forum with international participation, pp. 141-145. [In Russian]
6. Medialogy: How It Works. <https://tech.mlg.ru/> [In Russian]
7. Odintsov A. V. 2017. "The Sociology of Public Opinion and the Challenge of Big Data". Monitoring Public Opinion: Economic and Social Changes, no 3, pp. 30-43. [In Russian]
8. The National Security Strategy of the Russian Federation until 2020 (approved by Decree of the President of the Russian Federation of 12 May 2009 No 537). <http://kremlin.ru/supplement/424> [In Russian]
9. RF Federal Law of 28 December 2010 N 390-FZ "On Security". [http://www.consultant.ru/document/cons\\_doc\\_LAW\\_108546/](http://www.consultant.ru/document/cons_doc_LAW_108546/) [In Russian]
10. Boyd D., Crawford K. 2012. "Critical questions for BigData". Information, Communication and Society, vol. 15, no 5, pp. 662-679. DOI: 10.1080/1369118X.2012.678878
11. De Mauro A. 2016 "Formal definition of Big Data based on its essential features". Library Review, vol. 65, no 3, pp. 122-135. DOI: 10.1108/LR-06-2015-0061
12. De Mauro A., Greco M., Grimaldi M. 2014. "What is big data? A consensual definition and a review of key research topics". Proceedings of the 4th International Conference on Integrated Information. Madrid. DOI: 10.13140/2.1.2341.5048
13. Giddens A. 1987. The Nation State and Violence: Volume Two of a Contemporary Critique of Historical Materialism. University of California Press.
14. Giddens A. 1999. "Risk and Responsibility". The Modern Law Review, vol. 62, no 1, pp. 1-10. DOI: 10.1111/1468-2230.00188
15. Kitchin R., McArdle G. 2016. "What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets". In: Big Data & Society, pp. 1-10. DOI: 10.1177/2053951716631130
16. Lazer D., Kennedy R., King G., Vespignani A. 2014. "The parable of Google flu: traps in Big Data analysis". Science, 14 March, vol. 343, no 6176, pp. 1203-1205. DOI: 10.1126/science.1248506
17. Lupton D. 2015. "The thirteen Ps of Big Data". In: This Sociological Life: a blog by sociologist Deborah Lupton. <https://simplysociology.wordpress.com/2015/05/11/the-thirteen-ps-of-big-data/>