

Информатика сообществ и формирование социальных сетей

Данные как вызов

В.И. Тищенко

Федеральный исследовательский центр «Информатика и управление»
Российской академии наук», г. Москва, Россия

Аннотация. В статье рассмотрены проблемы анализа природы данных, возникающие на границе знания и бытия. Показана противоречивость представления науки о данных как совокупности технологий и алгоритмов, предназначенных для решения задач обработки больших массивов слабоструктурированных данных, прогнозирования нераспознаваемых ранее корреляций. Представлен этимологический и когнитивный анализ трансформации понятия «данные». Проанализированы концептуальные основания становления новой эпистемологической парадигмы отображения окружающего мира, позволяющей рассматривать по-новому проблему данных.

Ключевые слова: наука о данных, *data science*, эпистемология модели познания.

DOI: 10.14357/20790279240306 **EDN:** LHREBR

Введение

На современном этапе всесторонней цифровизации социального бытия и повсеместного экспоненциального роста количества данных «производство» («добыча»), как в обыденном, так и научном сознании, рассматривается как нечто очевидное, само собой разумеющееся. И, как следствие, для этого термина в сообществе специалистов сформировалось вполне рабочее и устоявшееся понимание [1]. *Данные* – это форма представления информации, фактов или понятий, зарегистрированных восприятий или измерений, пригодных для передачи, общения, интерпретации или обработки, в т. ч., с помощью цифровых устройств. Данные обычно рассматривают как мельчайшие единицы фактической информации,

которые могут быть использованы в качестве основы для расчетов, рассуждений или обсуждения. Данные могут варьироваться от абстрактных идей до конкретных измерений [2-5]. И, казалось бы, не совсем понятно, что может в настоящее время дать для развития информационных технологий когнитивный анализ, казалось бы, устоявшегося понятия «данные».

1. Многозначность взглядов на природу данных

Однако при всей кажущейся очевидности истолкования этого термина, значение данных как источника новых знаний, существующего на стыке знания и бытия и «возникающего» в моменты реги-

страции актов восприятия или измерений, не раскрывается в приведенных выше определениях и, с очевидностью, требует дополнительного рассмотрения [6]. К необходимости дополнительного осмысления природы данных, возможно, также подталкивает, то, что понятие «данные» нередко рассматривается как синоним понятия «информация», или элементов нового знания, и в то же время существует мнение, «данные больше, чем знания» [4].

Как ни странно, более глубокому рассмотрению этого понятия в немалой степени мешает проведенная в 2015 году стандартизация термина, предпринятая Международной организацией по стандартизации (ISO): «Данные – представляют поддающееся многократной интерпретации представление информации в формализованном виде, пригодном для передачи, связи, или обработки»¹ [ISO/IEC 2382:2015]².

Другими словами, это закрепило в общепринятом употреблении, представление о *данных* как наборе дискретных, или непрерывных значений, которые передают информацию, описывающую количество, качество, факты, статистику, другие базовые смысловые единицы или просто последовательности символов, которые в дальнейшем могут быть интерпретированы формально [9,10].

В академической среде стремление осмыслить природу данных как источника знаний тем не менее имеет определенную историю. В конце 80-х годов прошлого века это движение привело к построению т. н. «информационной иерархии» в форме четырехуровневой модели DIKW (данные, информация, знания, мудрость) соотношения данных и действительности [11]. Однако в дальнейшем интеллектуальный анализ данных, процессов их «создания» и сохранения, а главное аппаратное развитие информационных технологий показали, что модель DIKW предлагает сильно искаженную и упрощенную картину «извлечения» знания из данных. Попыткой уйти от схематизма модели DIKW и стало формирование направления, получившего название *наука о данных* (Data Science, англ.).

И, хотя начало исследований в области анализа данных обычно связывают с именами Д. Неймана и К. Пирсона, собственно Data Science как академическая наука начала развиваться лишь с 1996

г., даты создания CODATA (Комитета по данным для науки и техники), а также издания соответствующих научных журналов [12,13].

В основе этого направления находится интеллектуальный анализ данных и, соответственно, подготовка, отбор, очистка данных и адекватная интерпретация результатов [14]. Другими словами, Data Science исследует жизненный цикл *цифровых данных* — от появления до преобразования и представления в других областях знаний (Peter Naur).

Очевидно, что концептуальной основой Data Science является представление, что данные, и, соответственно их анализ, все же начинаются непосредственно со сбора и систематизации показаний, сгенерированных различными источниками в процессе преобразования потока независимых переменных. При этом данные отождествляются с зафиксированными наборами значений качественных или количественных переменных, которыми «обладают» и которыми «характеризуются» исследуемые процессы и события, существующие как бы независимо от наблюдателя. *Данные* – ведь, это то, что *дано*, *найденно*. И такое понимание представляется естественным, ведь слово *данные* в русском языке берет свое начало от праславянского слова **dati* – ‘*дар*’ (индоевропейское **dō*); ср. английское слово *data*, производное от латинского слова *datum* – *данная вещь* [15].

При этом основные усилия, по мнению IT-специалистов, затрачиваются при обработке данных на сбор, «очистку» и *систематизацию* полученных показателей. Ведь, пока *данные* не вписаны в тот или иной контекст, не структурированы, они не могут рассматриваться как исходный материал для аналитики, оставаясь как бы «сырыми» показаниями. Только в результате применения интеллектуальных алгоритмов из данных «извлекается» информация и получают «умные» («smart») данные [16]³. Но тогда анализ данных, по сути, рассматривается как способ нахождения адекватных алгоритмов, которые собственно и преобразуют, по мнению Data аналитиков, «сырые» зафиксированные переменные (которые ошибочно иногда определяют, как данные) в *данные*, пригодные для извлечения информации [17]. И следствием этого оказывается «естественный» вывод, что наука о данных – это разработка методов по обработке данных в условиях больших объемов.

В результате такого взгляда на проблему данных, вроде бы, исчезает необходимость обсуждать природу данных как таковых, их происхождение,

¹ «Data – a reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or communication, or processing» [7].

² В нашей стране разработан и утвержден аналог приведенного выше международного стандарта – ГОСТ 33707-2016 «Информационные технологии. Словарь»; согласно этому ГОСТу – «данные – предоставление информации в формальном виде, пригодном для передачи, интерпретации или обработки людьми, или компьютерами» [8, с. 21.].

³ Не удивительно, что здесь сразу вспоминается пирамидальная структура модели DIKW, где в основании лежит слой «данных», над которым располагается уменьшенный по сравнению с предыдущим слой «информация». И последующие слои – «знание» и «мудрость».

поскольку они представляют собой результаты «обработки» воздействий, зарегистрированных датчиками измерительных устройств. И, соответственно, наука о данных, рассматриваемая с этой точки зрения, обсуждает не столько собственно данные, сколько возможности их поиска, вычленения и агрегирования.

А между тем данные, сгенерированные разнообразными способами, не существуют сами по себе, и не появляются из ниоткуда. Ведь даже зарегистрированные показания (которые определяют, как источник данных) должны быть «порождены» регистрирующими устройствами, *выбранными* для описания (регистрации) «предстоящей» реальности согласно той или иной концептуальной модели исследуемой системы, будь она неосознанная или значимо формализована. Ведь анализ – это не ответ на вопрос, это «задание вопроса», нахождение связи между наблюдаемыми процессами. Поэтому наука о данных не является статистическим или компьютерным анализом обнаружения и выделения корреляций, которые могут свидетельствовать о наличии закономерностей и взаимосвязей в исследуемых процессах. И, соответственно, не «корреляции вытесняют каузальность», а каузальность порождает корреляции.

Данные без модели, которая «настраивает» регламент, процедуру, а самое главное, средства и методы сбора данных, или, по сути, определяет регистрирующие средства, приборы и датчики – всего лишь шум. Исследователь всегда воспринимает сведения, показания как определенный результат анализа реальности в ее целостности, рассмотрении в определенном «ментальном» ракурсе. Это означает, что в действительности никаких «сырых» данных (неструктурированных показателей), которые потом преобразуются в «smart» данные, не существует. Сама процедура регистрации, записи, описания значения переменных и есть способ представления и определения совокупности показаний, которые становятся данными уже на этапе, предшествующему их генерированию.

Так что же тогда такое *данные*? С одной стороны, они рассматриваются как нечто, созданное нашим восприятием, инструментами, или технологиями. В то же время, очевидно, что они «существуют» как бы вне нас со всем нашим инструментальным миром познавательной деятельности, как нечто *данное*. Чтобы разрешить это противоречие, полагаю, следует обратить внимание на то, как в языке, в семантической эволюции развития этого термина, отражается когнитивная «процедура» раскрытия самой природы термина *данные* как границы бытия и знания того, как реализуется переход от любых фактов к исчисляемым.

2. Когнитивная трансформация понятия данные

Диахронический анализ понятия «данные» следует начать с обращения к наиболее популярным (как бы «классическим») словарям. Ведь включение слов или фразеологизмов в словарь означает не только фиксацию статуса их как элемента общеупотребительной лексики, но как и процедуру их нормирования (нормализации). Поэтому крайне важно рассмотреть, как толкуют словари слово «данные» в том или ином языке. Так, согласно «*Оксфордскому словарю современного английского языка*»⁴ (1993), *данные* – это числовые и нечисловые значения характеристик кого-либо (чего-либо), операции с которыми проводит компьютер или какое-нибудь другое подобное устройство; данные представляют собой набор значений качественных или количественных переменных; отдельные фрагменты информации⁵. В «Новом словаре русского языка»⁶ (2000) термин *данные* определяются как «сведения, факты, характеризующие кого-либо, что-либо, необходимые для каких-либо выводов, решений» [19, с. 6]. Таким образом, в соответствие со словарным значением слово *данные* представляет собой набор значений качественных или количественных переменных.

Между тем подобное, как нередко считается, общепринятое истолкование *данных*, как основы вычисления при решении математических задач, в действительности возникло и сформировалось относительно недавно. Оно представляет собой результат семантической и когнитивной трансформации этого слова, полученной в процессе его исторического «освоения». Реконструкция происхождения слова *данные* показывает, что в русском языке оно восходит, как было упомянуто ранее, к этимологическому праславянскому гнезду **dati*, представляющему важный смысловой фрагмент в русской языковой картине мира.

Используя, как в быденном, так и в научном дискурсе это слово, мы, в то же время, не отдаем себе отчет в антропоцентрическом контексте его применения. В действительности, усваивая это слово в самом раннем возрасте посредством соци-

⁴ The Oxford Dictionary of Current English. by Della Thompson (Оксфордский словарь современного английского языка) содержит более 140 000 определений и статей, написанных для доступного освещения современного английского языка, источник всех наиболее часто используемых слов, значений, происхождения, употребления и правописания. [18, с. 4]

⁵ Перевод выполнен мною, ВТ.

⁶ «В толковом словаре Ефремовой насчитывается более 140 тысяч словарных статей толкование и значений большинства слов, примерно 11 тысяч значений слов являются отсылочными, что делает онлайн-словарь более удобным.» [19].

альной активности и обыденной речевой практики, мы воспринимаем его как отражение, некий естественный ментальный *образ* множества элементов (характеристик) реальности, которые существуют независимо от нас. Объективно! И в этом мы усматриваем основания реальности нашего познания действительности (ее освоение).

Тем самым, в нашей речи, и, соответственно, в сознании *данные* с самого начала овладения речевой деятельностью, как бы «противостоят» нам, нашей языковой личности (субъекту языка, языкового общения и речевой практики). Истоки этого противостояния заложены в исторической социальной деятельности, осваивающей и одновременно *созидающей* мир культуры, мир вещей, создаваемых и творимых *человеком* для человека. Перенимая все те специфически человеческие способы деятельности, которые предметно зафиксированы в формах создаваемых вещей, *осваиваясь* в мире этих вещей, мы усваиваем и опредмеченный в них общественно-человеческий разум с его логикой, выраженной в речевой деятельности, национальном языке (Ильенков Э. В.). Поэтому в ходе реконструкции отдельных элементов языка, словоупотребления мы раскрываем становление того или иного понятия как отражение истории практической жизни человека, формирования «следов» мышления. В поисках этих «следов», которые получили отражение в эволюции смыслов понятия *данные*, попробуем воспроизвести этимологическую реконструкцию данного слова.

Беря за точку семантического отсчета индоевропейский корень **dō-*, наш (русский) язык в своем развитии как бы порождает множество «семантико-словообразовательных центров **dati* – ‘*дать, давать, дарить*’, **danь* – ‘*дань, подать*’ – объект обязательной передачи в качестве платы’, **darь* – ‘*дар, подарок*’ – объект добровольной передачи в знак признательности, дружбы и установления мирных отношений’» [20]. Несмотря на очевидное многообразие элементов данного этимологического гнезда, мы видим, что в их смысловой основе лежит преобразованная семантика древнеиндоевропейского глагола *dā* – ‘*давать, дарить что-л.*’. Другими словами, значение понятия «данные» предстает перед нами как *нечто*, изначально независимо существующее и определяемое как бы без участия *того*, кому это нечто будет вручено, *подарено* (дано!). Но в то же время как нечто, которое уже исторически освоено и «вычленено» (выделено) из действительности, и, как результат, «присвоено» кем-то перед тем, как оно будет им передано.

Подобное семантическое «звучание» смысла слова *данные*, отмечается не только в русском или

родственных славянских языках, но и во множестве иных языков, которые восходят к индоевропейскому. Так, например, английское *data*, производное от латинского *datum* – ‘*данная вещь*’, берет свое начало от *dare* – ‘*давать*’. В этой связи, любопытно отметить, что древнеиндоевропейский корень **dō-*, в каком-то смысле, сохраняется в английском языке и в настоящее время в виде глагола неопределенной формы *do* – *делать*. Причем третья форма прошедшего времени этого глагола – *done* – и по произношению, и по правописанию коррелирует с русским словом *данные*.

В качестве еще одного примера раскрытия историчности смысла слова *данные* как отражения реальных («вещных») отношений в форме ‘*дать, давать, дарить*’ можно привести вывод, сделанный французским лингвистом Э. Бенвенистом. Так, он утверждал, что в хеттском языке корень **dō-* имел значение ‘*брать*’ и один из глагольных дериватов древнеиндоевропейского *dā* – *dāra* означал ‘*брать в жены, жениться*’. Из этого им был сделан вывод, что в индоевропейском языке, к которому восходит и хеттский язык, значения ‘*давать*’ и ‘*брать*’ синкретично соединены в одной словесной форме универсального обозначения обмена, занимающего важное место в праиндоевропейской картине мира [20, с. 82].

Важным аспектом рассмотрения возникновения и трансформации понятия *данные* представляется также анализ количественной формы использования этого слова. Она находится в прямой зависимости от ментальной среды того дискурса, контекста применения слова *данные*.

Так, если следовать «Энциклопедическому словарю Брокгауза и Ефрона», одно из сочинений Эвклида, представляющее продолжение его «Элементов», носило название «*Данные*» (δεδομένα). При этом Эвклид называет *данными* все то, что на основании теорем, доказанных в «Элементах» (предшествующего труда), непосредственно следует из условий задачи. Например, прямая, проводимая из точки по касательной к кругу, – есть *данная* по величине и положению [21, с. 133]. Другими словами, для Эвклида, заложившего основы математики, в содержании понятия *данные* отсутствовала количественная характеристика, но присутствовала деятельностьная (практическая) характеристика. Иначе говоря, *данные* – то, что мы изначально создали прежде, чем было начато математическое действие. Причины этого также обусловлены ментальными особенностями эпохи эллинизма.

Как видно из приведенного анализа, язык, речевая деятельность неразрывно связаны с историческими формами мышления, формами менталь-

ной активности. Эту зависимость не всегда удается выявить в явном виде, но в любом случае анализ семантики и грамматики словоупотребления показывает исторические особенности развития как обыденного, так и научного тезауруса.

В этой связи, кратко рассмотрим эволюцию использования существительных *data*, *datum* в английском языке, который можно рассмотреть, как языковую основу информатики. Слово *data* в значении, близком сегодняшнему, стало использоваться в начале 1600-х годов, будучи заимствованным из латинского языка, языка науки XVII-го и более ранних веков. Слово *data* произошло от латинского *datum* – «данная вещь» (см. латинское *dare* – «давать»). В английском языке оно применяется как во множественном, так и в единственном числе. Исчисляемое существительное *datum* означает «исходный факт», «заданная величина».

Первый случай использования термина «data» (*данные*) в англоязычном тексте был обнаружен в трактате 1630 г. Уильяма Баттена (William Batten), английского вице-адмирала, при составлении им навигационной таблицы, выполненной для вычисления местонахождения корабля по азимуту положения Солнца. У. Баттен называет цифры в своей таблице «данными», потому что, очевидно, он воспринимал их как данность положения корабля, поэтому сделанные записи не следует подвергать сомнению. Другими словами, по его мнению, *данные* описывают «аксиоматические» и, следовательно, неоспоримые утверждения или числа, *данные* («предоставленные») для расчета [22].

Несколько позже, в 1646 году, слово *data* (*datum*) в истолковании «куча данных» встречается в богословском трактате Генри Хаммонда (Henry Hammond) в форме утверждения – «не набор цифр, а список богословских положений, признанных истинными ради аргументации». В этом контексте использование при описании перечисленных богословских суждений слово «data» следует воспринимать как «данные», аналогично использованию У. Баттенем чисел уже как *данных* (вычисленных и внесенных) в его таблице. По мнению Генри Хаммонда слово Божие не подлежит ни сомнению, ни эмпирической проверке, его следует принимать как *данность*, как «аксиому». И понятие «данные» (терминологически как *data*) входит в английский язык как когнитивный «инструмент», сигнализирующий о результате освоения нового знания («информации»), которое следует воспринимать как данность, а не подвергать сомнению, будь то в области математики или теологии.

Осознание и «включение» в язык количественного значения терминов *datum* (*data*) произо-

шло позже в результате введения в сферу знания методов статического анализа демографических характеристик территорий Великобритании. В 1798 году Джон Синклер (Sir John Sinclair) в своем «Статистическом отчете о Шотландии» ввел в английский язык термин «*statistics*». Однако он считал, что «*statistics*» (статистика) – это собрание фактов, которые совершенно не обязательно представлены (и даже в основном не были представлены) в числовом виде [23].

И только с 1829 года слово *statistics* стало употребляться в английских документальных источниках в качестве «*собранных и классифицированных числовых данных*». Широкое распространение статистических методов в дальнейшем привело к тесной смысловой привязке термина *data* к термину *statistics* и, соответственно, пониманию *данных* как *числовых значений*. В результате с 1897 года в документальных источниках на английском языке слово *data* стало использоваться в новом значении – «*факты в числовом виде, собранные для дальнейшего использования*» [24].

В XX веке спектр значений слова *data* существенно расширяется и постепенно оформляется представление о *данных* как качественных и количественных переменных, отдельных фрагментах информации. В речевой практике появились и закрепились такие термины как «обработка данных» (*data processing*), «база данных» (*database*), «ввод данных» (*data entry*). И, наконец, в 1974 г. Питер Науэр (Peter Naur) публикует в Швеции и Соединенных Штатах книгу «Краткий обзор компьютерных методов» («Concise Survey of Computer Methods»). В этой монографии методы обработки данных соответствуют понятию *данные*, который был предложен Международной федерацией по обработке информации (IFIP). А именно: «Данные – это представление фактов или идей формализованным образом, способное передаваться или манипулироваться каким-либо способом»⁷. Исходя из этого, сам Питер Науэр определяет «науку, имеющую дело с данными»⁸, как Data Science. Пожалуй, это самое раннее сформулированное представление о Data science, как науке, работающей с данными после их «установления».

Описанная выше эволюция когнитивных смыслов понятия *данные* показывает, что анализ данных, определяемый таким образом и называемый в последствии, и как извлечение знаний, и как извлечение паттернов из данных, «добыча дан-

⁷ «A representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process» [3, c.9].

⁸ «Data science is the science of dealing with data, once they have been established, while the relation of data to what they represent is delegated to other fields and sciences» [3 c.13].

ных», и, наконец, сбор информации, в действительности отражает не столько процедуру выделения «полезных» данных из массивов зарегистрированных показателей в целом, сколько преобразование с помощью конкретных алгоритмов массивов независимых переменных, полученных при «участии» неких ментальных «усилий», в поток «исходящих» прогностических или информационных параметров. Между тем, трактовка понятия *данные*, как я считаю, зависит от того, как мы представляем процесс познания. Если в его основу мы помещаем классическую модель научного знания, согласно которой предметом исследования оказывается выделенная совокупность объектов, то наука о данных действительно имеет дело лишь с результатом их выделения. Но если мы рассматриваем науку о данных как процесс «освоения» и представления явлений и событий в когнитивном формате, то мы взаимодействуем с новым видением реальности. Возникающая в рамках науки о данных новая эпистемологическая установка подразумевает иной способ постановки исследовательских вопросов относительно производства знания, что ранее казалось не очень нужным, поскольку этот процесс рассматривался как естественное взаимодействие «наблюдателя» с реальностью. Тем самым феномен *данных*, особенно остро проявившийся в условиях всесторонней цифровизации социального бытия, ставит под сомнение устоявшуюся классическую схему представления об исследовательском процессе, как интерпретации, отображении «предстоящей» реальности.

3. Трансформация парадигмы познания

Необходимость формирования новой исследовательской парадигмы (познавательной деятельности в целом) в результате всесторонней цифровизации, получил отражение в утверждении – «большие данные» нуждаются в «большой теории». Крис Андерсон, еще в 2008 году, отмечал, что развитие инновационных технологий анализа данных указывает на новую стадию формирования знания, в то время как существующий научный метод знаменует «конец теории» [25]. Столь радикальный вывод был сделан им на основании того, что наука о данных, реализуя цели по извлечению знания из обрабатываемых данных, предстает как совокупность аналитических технологий выявления корреляций, отражающих наличие системных закономерностей и взаимосвязей в больших массивах данных.

Несмотря на привлекательность и, казалось бы, очевидность такого вывода, следует вновь отметить, что анализ данных начинается не с ис-

пользования статистических «технологий» или процедур обработки полученных данных, а с формулировки задач, которые мы хотим решить с помощью анализа данных. И сбор данных с последующей их обработкой с одной стороны – не более чем инструмент для реализации *правильно поставленной задачи*, а, с другой – некая модель, способ реализации нашей задачи, которая выражает проблему, требующую разрешения. При этом в основе, используемой/предложенной модели, лежит уже существующий в нашем совокупном знании о мире образ реальности, имманентно содержащий представление о возможности некой корреляции, которую мы намерены извлечь из «сформированного и заданного» нами набора значений, сгенерированных как «следствие» существующей корреляции. Поэтому реализуемая модель получения данных отражает, несмотря на кажущуюся непредвзятость, заложенные в нее цели, задачи, исследовательскую идеологию⁹.

В качестве разъясняющей аналогии любопытно привести описание процесса «возникновения» произведения искусства известного историка искусств Нильса Бюттнера (Nils Büttner). Создание любого произведения искусства, по его мнению, начинается с *'inventio'* (изобретения или обнаружения сущности), далее наступает стадия *'dispositio'* (этап выбора, оценки, классификации и структурирования полученного материала). И, наконец, завершается все *'elecutio'* (приданием формальным и материальным аспектам подходящей формы) [26]. Как мы видим, творческий процесс создания артефакта неотличимо совпадает с процедурой «формирования» некоего набора данных, представляющих фрагмент реальности. И это понятно, поскольку в основе этих различных форм человеческой деятельности, лежит процесс не просто освоения реальности, но воссоздание ментальных форм «присвоенной» реальности.

Предлагаемый взгляд на природу *данных* как источник знаний, «возникающий» на стыке знания и бытия в результате процесса «освоения» исследуемого фрагмента реальности, отражает формирование новой эпистемологической парадигмы, меняющей привычные процедуры интерпретации окружающего мира. Он знаменует новый подход к оценке произошедшего и прогнозу грядущего [27]. В отличие от предшествующих этапов развития науки о данных, формирующаяся парадигма характеризуется не только тем, что рассматривает современную эпоху в качестве цифровой эры анализа

⁹ И мы можем высказать «крамольное утверждение» – *данные* – это, по сути, модель представления осваиваемой в процессе осознания реальности во множестве ее особенностей и характеристик.

данных. И даже не тем, что разнообразные цифровые устройства или цифровые сети порождают посредством «заложенных» в них моделей обширные и динамичные потоки цифровых данных. Она ориентируется на то, что современное пространство социальной (человеческой) активности, преобразующей реальность и выраженной в совокупности форм создаваемых вещей, предстает как среда эффективных коммуникаций не только между людьми, но и различными предметами (Bruno Latour).

Датафикация социального пространства (*data everywhere*), в котором происходит становление и существование познаваемой реальности, приводит к тому, что мы рассматриваем реальный мир, с одной стороны, как множество данных, а с другой, как виртуальную среду коммуникационных взаимодействий различных сущностей (людей и предметов). В результате *данные* на современном этапе отражают цифровую реальность, которая оказывается порождением не только нашего знания (исследовательских практик), но и нашей социальной (человеческой) активности, «возникшей» в результате личностной социализации.

В контексте предлагаемой концепции *данные* рассматриваются как результат сложного процесса «разворачивания» процедуры структурирования предметного пространства, которое предстает перед исследователем как пространство социальной/познавательной коммуникации участников научного познания. И это пространство рассматривается в качестве конструкта, задающего концептуальный каркас исследуемого процесса научного познания. Соответственно, элементами пространства познания выступают не формы приращения знания (научное открытие, решение проблемы, фальсификация теории и т.п.), и не визуализированные формы этих приращений знания, документы, а *непосредственные* акты взаимодействия, которые и предстают в качестве анализируемых данных. Причем акты этого взаимодействия всегда конкретно историчны, и требуют дальнейшего исследования.

Заключение

Отсутствие достаточно ясного концептуального представления о том, как возникают данные на границе знания и бытия проявляется определенной многозначностью существующих определений понятия *данные*. Традиционно данные рассматриваются как первичные зарегистрированные воздействия, поступающие на органы чувств субъекта или датчики измерительного прибора. Однако определение данных как формы или способа

представления регистрируемых характеристик реальности не раскрывает их взаимосвязь с процедурами взаимодействия познающего субъекта с действительностью.

Вместе с тем, разработчики информационных технологий нередко рассматривают регистрируемые данные в качестве «сырых», не вписанных в тот или иной контекст. Только после соответствующего структурирования и выбора определенного формата представления *данные*, по их мнению, трансформируются в *исходный* «материал» для осуществления интеллектуальных операций, обеспечивающих осмысление и интерпретацию первичных зарегистрированных воздействий.

Между тем, представление о противостоянии *неструктурированных* и трансформированных («умных») данных не отражает реальную процедуру формирования знания в виде данных, «генерированных» различными источниками в процессе преобразования потока независимых переменных. Значения качественных или количественных переменных, которые характеризуют исследуемые процессы и события реальности, не могут быть получены независимо от теорий и моделей мышления исследователя. Он всегда привносит в исследование какую-либо концептуальную модель, будь она в имманентном виде, или формализована. Именно эта модель отражает совокупность исследовательских вопросов и способы их инструментальной постановки. И, соответственно, анализ начинается не с процесса обработки полученных данных, а с формулировки задачи, вопросов, которые мы хотим решить с помощью анализа данных.

Эпоха Data Science, проявляющаяся в повсеместном проникновении компьютерных технологий знаменует наступление нового эпистемологического этапа в познании реальности, так называемой «четвертой парадигмы» [The Fourth Paradigm]. Цифровизация социальной действительности, в которой происходит становление и существование познаваемой реальности, приводит к тому, что мы рассматриваем реальный мир, с одной стороны как множество данных, а с другой, как виртуальную среду коммуникационных взаимодействий не только между людьми, но и различными предметами (Bruno Latour). Результатом этого возникает цифровая реальность, которая, с одной стороны, есть порождение не только нашего знания (исследовательских практик), но и нашей социальной (человеческой) активности. И тем самым по-новому формулируется проблема данных как процедуры преодоления границы бытия и знания.

Литературы

1. *Gil Press*. A Very Short History of Data Science // Пост на сайте Forbes, updated Apr 14, 2022 [Электронный ресурс]. URL: <https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/> Дата обращения 17.06.2024
2. Журавлева Е.Ю. Эпистемический статус цифровых данных в современных научных исследованиях. Вопросы философии. № 2. 2012. С. 113–123.
3. *Naur Peter*. Concise Survey of Computer Methods. Lund, Sweden. Studentlitteratur. 1974.
4. *Tuomi Ilkka*. Data Is More Than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory // Journal of Management Information Systems. 2000. 6 (3). P. 103–117. DOI: 10.1080/07421222.1999.11518258. Дата обращения 17.06.2024.
5. «What are Data?» Australian Bureau of Statistics // [Электронный ресурс]. URL: <https://web.archive.org/web/20190419010315/http://abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+what+are+data>. Дата обращения 17.06.2024.
6. Смирнов Г.А. О месте данных в структуре познания // Труды ИСА РАН. 2019. Том 69 (3). С. 55–67. DOI: 10.14357/20790279190305
7. ISO/IEC 2382:2015 Information technology — Vocabulary. [Электронный ресурс]. URL: <https://webstore.iec.ch/publication/22380>. Дата обращения 18.06.2024.
8. ГОСТ 33707-2016 (ISO/IEC 2382:2015) Информационные технологии. СЛОВАРЬ. (ISO/IEC 2382:2015, MOD). М.: Стандартинформ. 2016.
9. OECD Glossary of Statistical Terms. OECD. 2008. DOI: [org/10.1787/9789264055087-en](https://doi.org/10.1787/9789264055087-en)
10. Data vs Information-Difference and Comparison [Diffen] [Электронный ресурс]. URL: https://www.diffen.com/difference/Data_vs_Information. Дата обращения 17.06.2024.
11. *Weinberger D*. The problem with the data-information-knowledge-wisdom hierarchy. Harvard Business Review. 2010. V.2/2. Feb 02. [Электронный ресурс]. URL: <https://hbr.org/2010/02/data-is-to-info-as-info-is-not>. Дата обращения 17.06.2024.
12. *Smith F. Jack*. Data Science as an Academic Discipline (англ.). Data Science Journal, 2006. 5(19). P. 163–164 // DOI:10.2481/dsj.5.163. Дата обращения 17.06.2024.
13. Тищенко В.И. Data science: новый этап сетевого анализа // Системный анализ и информационные технологии. САИТ-2019. Труды Восьмой международной конференции. 2019. М.: ФИЦ ИУ РАН. С. 370–377. DOI: 10.14357/SAIT2019048
14. *Fayyad U., Piatetsky-Shapiro G., Smyth P*. From Data Mining to Knowledge Discovery in Databases // AI Magazine. 1996; 17(3) [Электронный ресурс]. URL: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230>. Дата обращения 17.06.2024.
15. Пятаева Н.В. Этимологическое гнездо *dati в праславянском языке (реконструкция, словообразование, семантика) // Acta Universitatis Lodziensis. Folia Linguistica Rossica. 2017. (14). P. 113–123. DOI: <https://doi.org/10.18778/1731-8025>
16. *Goldman Jeremy*. Big Data is So. 2016. We Need Smart Data. Connecting the dots is more important than ever. // Блог Inc. Grow. 2017. March 21. // [Электронный ресурс]. URL: <https://www.inc.com/jeremy-goldman/big-data-is-so-2016-we-need-smart-data.html>. Дата обращения 24.05.2023.
17. *Breiman L*. Statistical Modeling: The Two Cultures Statistical Science 2001; 16(3), 199–231 [Электронный ресурс]. URL: <http://www2.math.uu.se/~thulin/mm/breiman.pdf>. Дата обращения 17.06.2024.
18. Oxford Dictionary of Current English / Edited by Della Thompson: 2nd Edition, Oxford University Press, 1993.
19. Ефремова Т.Ф. Новый словарь русского языка. Толково-словообразовательный. М.: Русский язык. 2000.
20. Пятаева Н.В., Булина Е.В., Ибатуллина Е.А. Дань, дар, подарки и духовные дары в динамике русской языковой картины мира // Актуальные проблемы филологии и педагогической лингвистики. 2018. 29(1). P. 103–112. DOI: 10.29025/2079–6021-2018-1(29)-103-112.
21. Эвклид. Энциклопедический словарь. Т. XL. С.-Петербург. 1904; 133.
22. *Hanlon Aaron R*. Data at the Dawn of the Anthropocene // The River Rail. 2019 [Электронный ресурс]. URL: [tps://brooklynrail.org/special/River_Rail_Colby/river-rail/Data-at-the-Dawn-of-the-Anthropocene](https://brooklynrail.org/special/River_Rail_Colby/river-rail/Data-at-the-Dawn-of-the-Anthropocene). Дата обращения 18.06.2024.
23. Сеница А.Л. Первое (старое) статистическое обозрение Шотландии как источник демографических данных о населении Шотландии конца XVIII в. // Вестник Томского государственного университета. 2018; 432. P. 159–169 [Электронный ресурс]. URL: <https://www.openrepository.ru/article?id=279663>. Дата обращения 19.06.2024.

24. Кузнецов С., Константинов А., Скворцов Н. Ценность ваших данных. М.: Альпина ПРО. 2022.
25. Anderson Chris. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete // Wired. June 23, 2008 [Электронный ресурс]. URL: <https://www.wired.com/2008/06/pb-theory>. Дата обращения 18.06.2024.
26. Бюттнер Н. «Таинственный Босх: кошмары Средневековья в картинах художника. М.: Эксмо. 2019.
27. The fourth paradigm: data-intensive scientific discovery, ed. by T. Hey, S. Tansley, K. Tolle, Kristin Tolle, Jim Gray. Microsoft Research, Redmond, Washington [Электронный ресурс]. URL: https://www.researchgate.net/publication/229529541_The_Fourth_Paradigm_Data-Intensive_Scientific_Discovery. Дата обращения 18.06.2024.

Тищенко Виктор Иванович. Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва, Россия. Заведующий отделом. Кандидат философских наук. Область научных интересов: системный анализ, информатика сообществ, виртуальные сообщества. E-mail: vtichenko@mail.ru

Data as a challenge

V. I. Tishchenko

Federal Research Center Informatics and Management” of the Russian Academy of Sciences”, Moscow, Russia

Abstract: The paper deals with the problem of analyzing the nature of data arising at the boundary of knowledge and being. It shows the contradictory view of data science as a set of technologies and algorithms designed to solve problems of processing large amounts of poorly structured data, predicting correlations unrecognised earlier. The etymological and cognitive analysis of the transformation of the concept “data” is presented. The conceptual foundations of the formation of a new epistemological paradigm for displaying the surrounding world, which allows us to consider the problem of data in a new way.

Keywords: *data science, epistemology of the cognitive model.*

DOI: 10.14357/20790279240306 **EDN:** LHREBR

References

1. Gil Press. A Very Short History of Data Science // Forbes, updated Apr 14, 2022; (In Russ). Available from: <https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/> [Accessed 17 June 2024].
2. Guravleva E.Ju. Epistemic status of digital data in modern scientific research. Voprosy filosofii, 2012; 2: 113—123 (In Russ).
3. Naur Peter. Concise Survey of Computer Methods. Lund, Sweden. Studentlitteratur. 1974.
4. Tuomi Ilkka. Data Is More Than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory // Journal of Management Information Systems. 2000. 6(3): 103—117. DOI: 10.1080/07421222.1999.11518258. [Accessed 17 June 2024].
5. What are Data? Australian Bureau of Statistics // Available from: <https://web.archive.org/web/20190419010315/http://abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+what+are+data>. [Accessed 17 June 2024].
6. Smirnov G.A. On the place of data in the structure of cognition. Trudy ISA RAN. 2019; 69 (3). 55-67. (In Russ). DOI: 10.14357/20790279190305.
7. ISO/IEC 2382:2015 Information technology — Vocabulary. Available from: <https://webstore.iec.ch/publication/22380>. [Accessed 17 June 2024].
8. GOST 33707-2016 (ISO/IEC 2382:2015) Information technologies. Vocabulary. (ISO/IEC 2382:2015, MOD). Moscow: Standartinform. 2016.
9. OECD Glossary of Statistical Terms. OECD. 2008. DOI: [org/10.1787/9789264055087-en](https://doi.org/10.1787/9789264055087-en)
10. Data vs Information-Difference and Comparison [Diffen Available from: https://www.diffen.com/difference/Data_vs_Information. [Accessed 17 June 2024].

11. *Weinberger D.* The problem with the data-information-knowledge-wisdom hierarchy. *Harvard Business Review*. 2010. V.2/2. Feb 02. Available from: <https://hbr.org/2010/02/data-is-to-info-as-info-is-not>. [Accessed 17 June 2024].
12. *Smith F. Jack.* Data Science as an Academic Discipline (англ.). *Data Science Journal*, 2006; 5(19). 163—164 // DOI:10.2481/dsj.5.163. [Accessed 17 June 2024].
13. *Tishchenko V.I.* Data science: new stage of network analysis // *Sistemnyi Analis. SAIT-2019. Trudy Vosmoi megdunarodnoi konfercii*. 2019; M.: FRC CSC RAS. 370-377. (In Russ) DOI: 10.14357/SAIT2019048
14. *Fayyad U., Piatetsky-Shapiro G., Smyth P.* From Data Mining to Knowledge Discovery in Databases. *AI Magazine* 1996; 17(3) // Available from: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230>. [Accessed 17 June 2024].
15. *Pyataeva N.V.* Etymological nest *dati in the Proto-Slavic language (reconstruction, word formation, semantics) // *Acta Universitatis Lodzianensis. Folia Linguistica Rossica*, 2017; (14), 113–123. (In Russ). DOI: org/10.18778/1731-8025
16. *Goldman Jeremy.* Big Data is so. 2016. We Need Smart Data. Connecting the dots is more important than ever. // *Blog Inc. Grow*. 2017; March 21. Available from: <https://www.inc.com/jeremy-goldman/big-data-is-so-2016-we-need-smart-data.html>. [Accessed 17 June 2024].
17. *Breiman L.* Statistical Modeling: The Two Cultures *Statistical Science* 2001; 16(3), 199-231 // <http://www2.math.uu.se/~thulin/mm/breiman.pdf>. [Accessed 17 June 2024].
18. *Oxford Dictionary of Current English*. Edited by Della Thompson: 2nd Edition, Oxford University Press, 1993.
19. *Efremova T.F.* New dictionary of the Russian language. Explanatory and word-formative. 2000; M.: Russk. Yaz. (in Russ).
20. *Pyataeva N.V., Bulina E.V., Ibatullina E.A.* Tribute, gift, gifts and spiritual gifts in the dynamics of the Russian language picture of the world. *Aktualnii problem filologii i pedagogicheskoi lingvistiki*. 2018; 29(1). 103-112. (In Russ) DOI: 10.29025/2079–6021-2018-1(29)-103-112.
21. *Euclid. S.* *Encyclopedicheskii Slovar*. XL. S. – Peterburg. 1904; 133.
22. *Hanlon Aaron R.* Data at the Dawn of the Anthropocene // *The River Rail*. 2019. Available from: https://brooklynrail.org/special/River_Rail_Colby/river-rail/Data-at-the-Dawn-of-the-Anthropocene. [Accessed 17 June 2024].
23. *Sinitisa A.L.* The first (old) statistical review of Scotland as a source of demographic data on the population of Scotland at the end of the 18th century. *Vestnik Tomskogo gosudarstvennogo universiteta*. 2018; 432. 159-169 (In Russ) Available from: <https://www.openrepository.ru/article?id=279663>. [Accessed 18 June 2024].
24. *Kuznetsov S., Konstantinov A., Skvortsov N.* Value of your data. M.: Alpina PRO. 2022.
25. *Anderson Chris.* The End of Theory: The Data Deluge Makes the Scientific Method Obsolete // *Wired*. June 23. 2008; Available from: <https://www.wired.com/2008/06/pb-theory>. [Accessed 18 June 2024].
26. *Byutner N.* Mysterious Bosch: nightmares of the Middle Ages in the artist's paintings. 2019. M.: EKSMO. (In Russ)
27. *The fourth paradigm: data-intensive scientific discovery*, ed. by T. Hey, S. Tansley, K. Tolle, Kristin Tolle, Jim Gray. Microsoft Research, Redmond, Washington Available from: https://www.researchgate.net/publication/229529541_The_Fourth_Paradigm_Data-Intensive_Scientific_Discovery. [Accessed 19 June 2024].

Tishchenko Viktor I. Federal Research Center “Informatics and Management” of the Russian Academy of Sciences”, Moscow, Russia. Head of Department. Ph.D. Area of scientific interests: system analysis, community informatics, virtual communities. E-mail: vtichenko@mail.ru