ИНТЕРНЕТ-ДАННЫЕ И ИХ РОЛЬ В СОВРЕМЕННОЙ НАУЧНОЙ ДЕЯТЕЛЬНОСТИ

Е.Ю. Журавлева

За последнее время в мире осуществлено множество исследований, целью которых является выявление суммарного объема цифровых данных. Исследовательская компания IDC опубликовала прогноз (The Expanding Digital Universe: A Forecast of Worldwide Information Growth Through 2010 www.emc.com/about/destination/digital_universe), согласно которому к 2011 г. общие объемы данных, хранимые во всем мире на всех существующих цифровых носителях, превысят 1800 экзабайт, что в 10 раз больше, чем в 2006 г., что объясняется ростом популярности цифровых медианосителей и объемов графических, аудио- и видео-файлов в глобальной сети Интернет. По своей структуре существующие файлы различны (от 6-гигабайтные образов DVD-дисков до мета-файлов RFID-меток, размер которых не превышает 128 байт). Было выяснено, что пользователями активно эксплуатируется менее половины из всех хранимых данных, остальные файлы - это, так называемые «цифровые тени», к которым относятся файлы интернет-кешей в браузерах, журнальные файлы на серверах, данные об уже совершенных транзакциях, истории веб-поиска и т. п. Эксперты компании Cisco (http://www. cisco.com /) также предсказывают увеличение объемов цифровых данных в сети Интернет, но приводят другие цифры: количество переданных в Интернете данных будет ежегодно расти примерно на 46%, что в результате приведет к достижению объема в 522 экзабайт.

Возрастающий объем цифровых данных и изменение структуры их использования начинают играть все более значимую роль в современном научном познании. Научно-исследовательскую деятельность в настоящее время все больше следует рассматривать в прямой зависимости от эффективного доступа к общим цифровым научным данным и к современным информационным инструментам, которые позволяют осуществлять хранение, поиск, визуализацию и высокий уровень анализа данных. Интернет стал популярным средством для сбора данных из-за его способности получить доступ к миллионам пользователей, возможности исследования массива данных и удобства технологических процедур.

Понятие «данные» обычно рассматривается в контексте развития информационно-коммуникационных технологий, его соотношения с информацией и знанием. Современные информационные технологии предоставляют исследователю мощный аппарат для работы с данными, причем с их вычислениями, но не с операциями над ними.

Научных определений данных по сравнению с определениями информации не так много. Рассмотрим самые распространенные, данные - это сведения, служащие для какого-либо вывода и возможного решения. Они могут храниться, передаваться (в форме баз данных), но не могут уже выступать в качестве информации [1]. С. А. Нехаев, И.Л. Андреев, Н.В. Кривошеин, Я.С. Яскевич [2] наоборот предполагают, что данные это информация, представленная в формализованном виде, пригодном для автоматизированной обработки. На взаимосвязь данных и информации указывает определение - информация это данные, привязанные к конкретной модели или наделенные семантической структурой [3].

Р. С. Гиляровский расширяет определение данных, под которыми понимает факты, идеи, сведения, представленные в знаковой (символьной) форме, позволяющей производить их передачу, обработку и интерпретацию [4].

В экономике знаний циркулирует определение, что данные это код и с ним возможны следующие операции: кодирование, вычисление, передача, накопление и порождение новых комбинаций.

По версии Р. М. Юсупова и В. П. Котенко данные являются одной из основных форм представления знаний в информатике, наряду с базами данных, базами знаний, текстами, гипертекстами, когнитивными информационными моделями. Данные можно рассматривать как представление понятий и фактов для интерпретации (истолкования смысла) [5].

Данные в Интернете имеют уникальные особенности и исследователи Л. Флориди (L. Floridi), М. Хейм (M. Hime), А. Биверс (A. Bifers) подчеркивая эти особенности, вводят понятие «Интернет-данные». В целом, Интернет можно рассматривать как источник новых данных, и в этом случае будет уместно ввести понятие «Интернет-данные», и как новый источник для уже имеющихся данных (обычно такие данные называются цифровыми).

Практическое освоение понятия «Интернет-данные» происходит в контексте создания семантического Веба, поисковых систем, экспертных систем, систем виртуализации, интеллектуальных агентов, первого и второго поколения ГРИД. Познавательное содержание «Интернет-данных» опосредованно раскрывается в разделах эпистемологии: информационной, социальной, эволюционной, кибернетической и т. д.

Интернет-данные (internet-date, e-date) - это представление фактов, результатов экспериментов и идей в формализованном виде, пригодном для передачи и обработки в информационном процессе, а также выделенная из системы информация или часть программы, совокупность значений определенных ячеек памяти, преобразование которых осуществляет код.

Интернет-данные имеют следующую специфику: гетерогенны, наделены высокой автономией, имеют неограниченный объем и соответствуют определенным техническим стандартам и нормам (форма представления, канал, трафик). Цифровые данные имеют короткий диапазон существования, так они должны быть переданы для обработки или хранения, иначе они в дальнейшем становятся нечитаемыми.

По степени структурированности данные в сети Интернет можно разделить на структурированные, полуструктурированные и неструктурированные. Критерием деления является соответствие определенному формату. В структурированных данных отражаются отдельные факты предметной области (это основная форма представления данных в системах управления базами данных, СУБД). Именно структурированные данные имеют наибольший интерес, так как они связаны с другими данными и представляют информационную ценность с точки зрения количества информации.

Полуструктурированные данные это данные, которые имеют характеристики схем и метаданных. Стандартом представления полуструктурированных данных является XML [6]. Понятие «метаданные» многозначно, оно может означать информацию о данных, или структурированные данные, представляющие собой характеристики описываемых сущностей для целей их идентификации, поиска, оценки, управления ими, а также данные из более общей формальной системы, описывающей заданную систему данных. Метаданные имеют большое значение в Интернете по причине необходимости обеспечения поиска полезной информации среди огромного количества доступной. Метаданные, созданные вручную имеют большую ценность, поскольку это гарантирует их осмысленность.

Под неструктурированными данными понимаются произвольные по форме текстовые документы (тексты естественного языка), электронные таблицы, сообщения электронной почты, графика, музыка, видео и т. д. Эта форма представления данных широко используется, например, в Интернет-технологиях, а сами данные предоставляются пользователю в виде отклика поисковыми системами. По современным оценкам более 95% цифровой среды состоит из неструктурированных данных. Многие исследователи работают в Интернете именно с неструктурированными и слабо структурированными данными, совокупность которых называют пространствами данных.

Интернет-данные по мере обработки можно подразделить на первичные, вторичные и сводные. К категории первичных данных относят слабо упорядоченный набор фактов, характеризующих определенное явление, это могут быть данные полученные в виде интервью, документальных записей событий, контент-анализа прессы. Вторичные данные являются результатом определенного логического осмысления фактов со стороны непосредственных участников или внешних наблюдателей, этот тип данных может быть получен из работ других авторов или опросов экспертов. Сводные данные это сведения, объединенные в большие специальные группы и отличающиеся друг от друга разнообразной формой. Существует несколько типов основных типов сводных данных с различной степенью валидности для аналитических заключений: данные переписи, статистика, материалы тематических публикаций, событийная информация и экспертные оценки [7].

Дж. Бэнфильд и Дж. Слэмко (Benfield J. A., Szlemko W. J., 2006) считают, что Интернет сравнительно мало используются для первичного сбора данных во многих научно-исследовательских областях [8]. Например, исследователи в области социальных наук еще только начали реагировать на появление Интернета, это показал экспертных обзор 494 статей с ключевым словом «Интернет-исследования» опубликованных в крупных социологических журналах с 1996 по 2006 гг. Интернет рассматривается как богатый источник для литературы и вторичных данных исследований по социальным наукам. Сравнительно недавно использование Интернет для первичного сбора данных требовало от исследователя знаний основ программирования. Но созданные технологические решения и услуги позволяют ученому проводить сетевые исследования не изучая программирование. Поиск в библиографической базе данных Web Science® показывает, что число публикаций в течение шестилетнего периода 2000-2005 гг., с использованием ключевого слова «Интернет-исследования», составляет 128, которое на 312% превышает соответствующий показатель за шесть лет период до 2000 г., т.е.,1994-1999 гг. Аналогичные результаты отмечались при поиске фраз «сбор данных в Интернете» (325%), «веб-исследования» (333%) и «электронный сбор данных» (327 %). Конечно, эти впечатляющие показатели основаны на низких базовых цифрах, но использование Интернет-данных в научных исследованиях все еще остается довольно ограниченным.

Над Интернет-данными возможны операции создания, генерации, обработки, хранения и распределения. Новое научное направление, целью возникновения которого стали вычисления над данными большого объема, получило название Data-Intensive Computing. Проблемное поле Data-Intensive Computing включает в себя две области: управление и обработка экспоненциально возрастающими объемами данных, поступающих в реальном времени в виде потоков данных от приборов, или генерирующихся в ходе имитационного моделирования; и сокращение времени анализа данных для возможности своевременного принятия решений исследователями.

Хранение промежуточных и итоговых данных Интернет-исследований тоже является сравнительно новой научной областью и ее понятийный аппарат не являются достаточно устоявшимся. П. Лорд, А. Макдональд, Л. Лаон, Д. Джаретто (Lord P., Macdonald A.,Lyon L., Giaretta D.) предлагают использовать следующие рабочие определения [9]:

- *хранение* - деятельность по управлению и содействию использования данных. Для динамических данных это может означать непрерывное обновление.
- *архивирование* - деятельность, которая гарантирует, что данные правильно выбраны, хранятся и что их логическая и физическая неприкосновенность сохраняется с течением времени.
- *сохранение* - деятельность в архиве, в котором по конкретным пунктам данные поддерживаются на протяжении всего времени с тем, что они все еще могут быть доступны и понятны путем внесения изменений.

Особое значение для понимания Интернет-данных приобретают базы данных, банки данных и хранилища данных. База данных (database) совокупность данных, организованных по определенным правилам, предусматривающим общие принципы описания, хранения и манипулирования, независимая от прикладных программ. База данных является информационной моделью предметной области. Обращение к базам данных осуществляется с помощью системы управления базами данных (СУБД). Использование в научной деятельности баз данных изменяет процесс получения знаний. Базы данных применяются как крупномасштабное средство коммуникации исследователей и, лишь в меньшей степени в качестве инструмента познания, замечает К. Хине [10].

В свою очередь Банк данных (databank) это автоматизированная информационная система централизованного хранения и коллективного использования данных. В его состав входят одна или несколько баз данных, справочник баз данных, СУБД, а также библиотеки запросов и прикладных программ [2].

Для аналитической обработки предназначено собрание данных, отличающихся предметной ориентированностью, интегрированностью, поддержкой хронологии, неизменяемостью, которое носит название хранилище данных (data warehouse). В настоящее время в связи с возрастанием количества научно-исследовательских данных приобретают сетевые услуги хранения научных данных (например, esnips.com).

Стремительное развитие глобальной информационной сети Интернет ведет к изменению фундаментальных парадигм обработки данных, которые можно охарактеризовать как переход к распределенным ресурсам и создание инфраструктуры для свободного доступа к ним. Организация инфраструктуры для свободного доступа к данным требует создания служб: публикации данных, поддержки их аутентичности и качества; специальных поисковых систем обнаружения информации и анализа распределенных данных.

Как отмечает З. А. Сокулер обобщение массивов данных в реальном времени для отслеживания происходящих процессов – в экономике, погоде, эпидемиях и т. п., создает возможность социально-экономических, политических и других типов исследований [11].

В настоящее время осуществляется несколько высокотехнологических проектов по использованию Интернет-данных в научной деятельности. Проект Корнел (Cornell) создан как научно-исследовательская лаборатория для исследований в области социальных наук на основе Интернет. Базы данных включают в себя архив из 40 миллиардов веб-страниц и 200 терабайты социальных исследований. Некоммерческий Интернет-архив снимков особенностей всех данных в World Wide Web, собирается каждые два месяца в течение 10 лет с 1996 по 2005 гг. В рамках проекта можно скопировать и перенастроить большую часть этого массива данных как реляционную базу данных, которая может быть использована для исследований в области социальных и информационных сетей [12]. Среди других сфер исследований можно назвать поиск данных, изучение сообществ, в том числе на MySpace, Facebook и Живого журнала. Исследователь М. Меис (M. Macy) считает, что «сетевые взаимодействия оставляют цифровые следы, что создает беспрецедентную возможность для исследования социальной жизни на реляционной уровне». Все это должно способствовать исследованиям в анализе социальных сетей и в разработке дополнительных инструментов для дальнейших исследований приложений социальных наук [13].

Общество физиков по изучению элементарных частиц планирует серию экспериментов, которые нужно выполнять на «Large Hadron Collider» (LHC) созданном в CERN (Женева). К эксперименту на LHC привлечены свыше 100 учреждений и свыше 1000 физиков из Европы, США и Японии. В ходе ряда экспериментов обширная сумма сгенерированных данных (10 Petabytes) должна быть переработана и распространенна для дальнейшего анализа всеми участниками консорциума[14].

DSpace (DSpace Project, www.dspace.org) это проект по разработке открытой платформы цифрового архива реализованный совместно компаниями Hewlett-Packard и MTI (Массачусетским Технологическим Институтом). Система предназначена для хранения цифровой структурированной и неструктурированной информации. DSpace позволяет захватывать, управлять, индексировать, хранить и предоставлять доступ к архивам научной информации, включая книги, монографии, диссертации, карты, изображения, аудио/видео информации, базы данных, веб-страницы и т.д. Система используется как цифровой архив более чем 200 академическими институтами.

Р. Акланд (Ackland R.) создал виртуальную обсерваторию по изучению сети Интернет (VOSON), которая является системой, основанной на Интернет-технологиях по добыче и визуализации данных. Проект VOSON использует новейшие электронные исследования – киберинфраструктуру и является примером в области электронной социальной науки[15].

Интернет-данные могут быть использованы при изучении самого Интернета как сложной самоорганизующейся системы. Разработка анализа показателей сети является прогрессирующей областью исследований, которая включает в себя Киберметрику, Вебметрику. Официальным образованием Киберметрики (cybermetrics) считается год основания электронного журнала Киберметрика[16] в 1997 г. Европейский союз признал важность Вебметрики путем финансирования двух крупных проектов «Веб показатели для науки, технологии и инновационной научно-исследовательской деятельности» [17] и «Европейские показатели, киберпространство и научно-технологическая-экономическая система[18]».

Основной целью Вебметрики (Webometrics) как науки является измерение Мировой Сети, для того чтобы получить знания о числе и типах гиперлинков и структуре Интернет. Согласно Вьюнбону и Ингвешену (Bjrneborn, Ingwersen, 2004), «Вебметрика это исследование количественных аспектов конструкции и использования информационных ресурсов, структур и технологий Web, с использованием библиометрических и инфометрических методов». Пограничными по интересам научными областями Вебметрики считается Библиометрика, Инфометрика, Наукометрика, Виртуальная этнография и сетевой анализ. Интернет-данные об измерениях обеспечивает основу для функционирования и планирования сетей, составляющих Интернет, и являются необходимым компонентом в области научных исследований для анализа, моделирования и эмуляции.

Одним из индикаторов изучения Интернета является «Показатель Влияния Web» (WIF) введенный Ингвешеном (Ingwersen, 1998). Показатель WIF может быть определен как номер страниц веб на веб-сайте, получающем ссылки из другое места веб, поделенные на номер веб-страниц опубликованного в сайте, который доступен поисковику. Согласно Нерузи (Noruzi, 2006), показатель WIF высчитывается, для того чтобы прокладывать навигацию только в пределах вебсферы страны, используя один язык и единственную подчиненную область.

Интернет-исследования предполагают новые формы и уровни исследований[19]. На различных этапах исследований образуется огромное количество данных, поэтому в связи с понятием «данные» целесообразно ввести понятие «поток данных» (data deluge).

Понятие «поток данных» возникло в науке одновременно с развитием высокотехнологичных инструментальных исследований в физике, астрономии, химии, молекулярной биологии и в настоящее время все больше распространяется в другие сферы. Внедрение «сенсорных сетей» является в настоящее время важными технологиями для наук по изучению окружающей среды. Интернет-коммуникации предоставляют многочисленные «следы» человеческой деятельности для исследователей, занимающимися общественными дисциплинами. Педагоги регистрируют взаимодействия с симуляцией экспериментов, совместными инструментальными средствами и внедрениями оценок. Ученые-гуманитарии увлечены глубинным анализом текстов и моделированием сообществ [20].

Данные из широкого ряда новых источников должны быть записаны как метаданные, заархивированы и сохранены, чтобы как данные, так и программы могли использоваться и воспроизводиться в будущем. Интернет-исследователям необходимы распределенные источники разнообразных типов данных и ресурсов, чтобы анализировать или представлять себе путь исследования. Т. Хей и А. Трефэзен (Hey T., Trefethen A., 2006) рассуждают о создании программного обеспечения (ПО), которое будет посредником между Интернет-данными и технологиями их обработки [21]. В качестве примера такого типа ПО выделяют сервисы SRB[22] и Globus middleware[23].

Дж. Д. Майерс (J. D. Myers) в свою очередь рассуждает о потоке метаданных (Metadata Deluge) [24], который необратимо образуется при автоматизации процессов хранения и обработки потоков данных. Обеспечение развития системы сохранности данных, их хранения является чрезвычайно трудоемким. Если в ближайшее время возникнет поток метаданных, то он глубоко повлияет на роль инфраструктуры хранения данных.

Таким образом, современная наука становится все более зависимой от генерации и повторного использования огромных массивов данных. Массив мультидисциплинарных данных растет быстрыми темпами и в беспрецедентных масштабах, а существующий «поток данных» может стать неотъемлемым компонентом современной научной инфраструктуры производства знания. На фоне этого растущего энтузиазма не нужно забывать многочисленные недостатки, которые угрожают качеству цифровых данных, их читабельности, полезности и способам хранения. Работа с цифровыми данными потребует мобилизации ученых для их экспертизы. В данном случае могут возникнуть проблемы в режимах управления системами данных и их продуктами, в контроле качеством цифровых данных и в вопросах авторского права. Передача, обработка и хранение огромного объема Интернет-данных потребует построения новых научных моделей, взамен традиционным.

ЛИТЕРАТУРА:

1. Соколова И.В. Социальная информатика и социология: проблемы и перспективы взаимосвязи - М.: Союз, 1999. - 228с.
2. Нехаев С. А., Андреев И.Л., Кривошеин Н.В., Яскевич Я.С. Словарь прикладной интернетики. Сетевой холдинг WEB PLAN Group. [WWW document] URL http://www.webplan.ru/hold/r15-4.shtml

3. Федотов А. М., Барахнин В. Б. Ресурсы Интернет как объект научного исследования http://www.csr.spbu.ru/pub/RFBR_publications/articles/computer%20science/2007/resursy'_internet_07_inf.pdf
4. Гиляровский Р. С. Основы информатики. М.: Изд-во МГУ. 1998.
5. История информатики и философия информационной реальности: Учебное пособие для вузов / Под ред. чл.-корр. РАН Р. М. Юсупова, проф. В. П. Котенко. – М.: Академический Проект, 2007. С. 194
6. Extensible Markup Language, расширенный язык гипертекстовой разметки
7. Боришполец К. П. Методы политических исследований: Учебное пособие для студентов вузов. М: Аспект Пресс. – 2005. – с. 36.
8. Benfield J. A., Szlemko W. J. Internet-Based Data Collection: Promises and Realities // Journal of Research Practice Volume 2, Issue 2, Article D1, 2006 http://jrp.icaap.org/index.php/jrp/article/view/30/51
9. Lord P., Macdonald A.,Lyon L., Giaretta D. From Data Deluge to Data Curation http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/150.pdf
10. Hine C. Databases as scientific instruments and their role in the ordering of scientific work // Social Studies of Science, 2006. - 36(2), 269-298.
11. Сокулер З. А. Детрансцендентализированный субъект в обществе знаний. http://kosilova.textdriven.com/sokuler/print
12. Сайт проекта Cornell http://www.news.cornell.edu
13. http://www.news.cornell.edu/stories/May06/ISS.networktheme.dea.html
14. GridPP: http://www.gridpp.ac.uk/, Griphyn: http://www.griphyn.org/, The Particle Physics DataGrid: http://www.ppdg.net/
15. VOSON http://voson.anu.edu.au
16. Cybermetrics , http://www.cindoc.csic.es/cybermetrics/cybermetrics.htm
17. WISER, Web Indicators for Science, Technology & Innovation Research, http://www.wiserweb.org/
18. EICSTES, European Indicators, Cyberspace and the Science-Technology-Economy System, http://www.eicstes.org/
19. Подробнее: Журавлева Е. Ю. Internet-research: сущность, структура и методы. Вологда: Легия. – 2009. – 224 с.
20. Borgman C. L. Will the Data Deluge Improve or Impair the Quality of Scholarship?
21. http://ora.ouls.ox.ac.uk/objects/uuid%3A64aa6f39-7e81-4d42-a008-ee2d7524bd67
22. Hey T., Trefethen A. The Data Deluge: An e-Science Perspective
23. http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/research/esci/datadeluge.pdf
24. The Storage Resource Broker, http://www.npaci.edu/DICE/SRB
25. The Globus project, http://www.globus.org
26. Myers J. D. The Coming Metadata Deluge http://www.arl.org/bm~doc/metadata.pdf