УДК 025.7/9+025.5

А. И. Земсков

Data Curation – хранение научных данных и обслуживание ими – новое направление деятельности библиотек

Цель статьи – кратко охарактеризовать новое направление деятельности вузовских, научных и научно-технических библиотек – хранение и обслуживание научными данными.

Использованы материалы трёх докладов крупных европейских специалистов, а также – Билефельдской конференции и Всемирного конгресса ИФЛА, состоявшихся в 2012 г.

Ключевые слова: информационное обслуживание, университетские библиотеки, научно-технические библиотеки, научные данные, хранение, доступ, навигация, электронные ресурсы, библиотеки данных, Билефельдская конференция 2012 г., Генеральная конференция ИФЛА 2012 г.

Термин *data curation* в последнее время стал часто использоваться в профессиональной (библиотечной) литературе, особенно, когда речь идёт о будущем вузовских, научных и научно-технических библиотек. Обсуждению готовности библиотек и библиотекарей к реализации технологий *data curation* посвящено множество специальных семинаров, выступлений на конференциях и публикаций.

В русскоязычной профессиональной лексике эквивалентный термин пока не установился. Поскольку глагол *to cure* (англ.) имеет несколько значений – 1) хранить, заботиться (отсюда – куратор выставки в музее, куратор проекта), 2) лечить и выхаживать после болезни, 3) обеспечивать сохранность пищевых продуктов путём, например, засаливания, маринования или копчения, – я перевожу этот термин, используя словосочетание, вынесенное в заглавие статьи – *хранение научных данных и обслуживание ими*. Однако сам ощущаю неполноту, некоторую корявость перевода, и если кому-либо из уважаемых читателей удастся предложить более складный термин, буду благодарен.

На эту технологию я обратил внимание ещё на Билефельдской конференции в 2009 г.: тогда тема хранения и повторного использования научных данных прозвучала в выступлении Герберта Ван де Сомпеля (Herbert Van de Sompel) – выдающегося специалиста по информационному обслуживанию учёных, сотрудника Лос-Аламосской научной лаборатории США (http://conference.ub.uni-bielefeld.de/2009/programme/). Напомню, что именно в Лос-Аламосе в 1991 г. по инициативе профессора Поля Гинспарга началось движение за открытый доступ, а Сомпель с коллегами создал систему протоколов открытого доступа – OAI-PMH.

Отметим, что технология обслуживания научными данными, *data curation*, часто рассматривается в комплексе с технологией связанных данных – *linked data*, т.е. осмысленного использования соотношений между различными наборами данных.

Введение в тематику *data curation* будет неполным без знакомства с тремя ключевыми докладами, её освещающими.

Первый доклад – «Оседлав волну…». По поручению Генеральной дирекции по информационному обществу и медиа Европейского Союза, группа высококвалифицированных экспертов: председатель – Джон Вуд (*John Wood*), генеральный секретарь Ассоциации университетов стран Британского содружества; в составе группы такие известные специалисты, как Ахим Бахем (*Achim Bachem*), председатель совета директоров Исследовательского центра Юлих, Герберт ван дер Сомпель, Йенс Виген (*Jens Vigen*), директор библиотеки ЦЕРН, и другие, в октябре 2010 г. подготовила доклад «Оседлав волну. Каким образом Европа может получить пользу от нарастающей волны научных данных» (*Riding the wave. How Europe can gain from the rising tide of scientific data.The High Level Expert group on Scientific Data*; http://www.cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf).

В докладе отмечено, что изобилие научных данных меняет природу научного исследования, позволяя учёным, которые располагают самыми разными базовыми возможностями, работать с одними и теми же

наборами данных, тем самым потенциально помогая решению крупных общественных проблем. Группа представила своё видение на перспективу до 2030 г. и соответствующие рекомендации.

По мнению Группы, к 2030 г. все участвующие в работе с научными данными – от учёных до правительств и широкой публики – поймут критическую важность сохранения и совместного использования научных данных на базе созданной к тому времени мировой инфраструктуры.

Второй доклад – «Структурный анализ. Взгляд изнутри». Очень солидно, с массой аналитических выкладок выполнен проект «Структурный анализ. Взгляд изнутри. Проблемы обеспечения постоянного доступа к результатам европейских научных исследований» (PARSE.Insight: INSIGHT into issues of Permanent Access to the Records of Science in Europe (deliverable 3.4 of PARSE.Insight). Его авторы – Том Куйперс (*Tom Kuipers*) и Джефри Ван дер Хойвен (*Jeffreyvan der Hoeven*) – сотрудники Королевской (национальной) библиотеки Нидерландов (цитируется по опубликованной версии 3.4).

Обилие электронных ресурсов составляет основу интеллектуального капитала европейских исследований, суть которых – обнаружение информации в этих ресурсах и предоставление новому поколению исследователей возможности «встать на плечи гигантов». Эти электронные ресурсы должны продолжать своё существование и оставаться доступными для поиска. Повторное использование научных данных (например пользователями, работающими в иной научной области) должно осуществляться немедленно после получения результатов – иначе этого не случится в течение длительного времени. Существует весьма реальная угроза, что большинство научных данных и документов, имеющихся сегодня, могут оказаться утраченными для будущих поколений, если не будут приняты меры по обеспечению постоянного доступа к ним.

Проект, выполненный за два года (2009–2010 гг.), сфокусирован на изучении потребностей формирования инфраструктуры, необходимой для поддержания постоянного и долговременного доступа и навигации по массиву научных данных.

Для проведения исследования были выделены четыре группы участников: 1 – учёные, 2 – организации и сотрудники, работающие с научными данными (в архивах, библиотеках и т.п.), 3 – издатели, 4 – финансирующие организации.

Технология исследования включала в себя: изучение литературы, обзоры, интервью, рассмотрение конкретных примеров. На вопросы получено 1 840 ответов: 1 389 – от учёных, 273 – от специалистов по обработке данных, 178 – от издателей; количество ответов финансирующих организаций пока невелико и недостаточно для анализа. «География» ответивших на вопросы представлена на рис. 1.

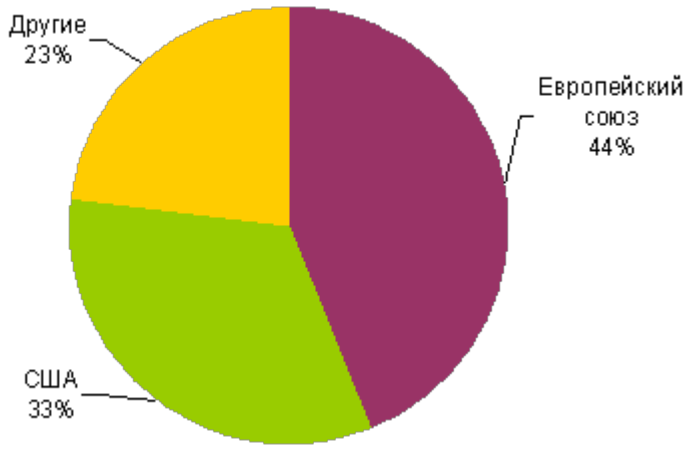

Рис. 1. Распределение – «география» – ответов, полученных в ходе исследования

Общее количество учёных в Европе – 1,33 млн. Для достижения статистической доверительной погрешности ниже 5% требовалось опросить не менее 385 человек.

По сравнению с представителями других научных дисциплин, наибольшее количество ответов поступило от физиков – 33%; специалисты с опытом работы более 20 лет оказались активнее своих молодых коллег.

Рассмотрим основные результаты и выводы исследования.

Позиция учёных. Причины, по которым следует заботиться о сохранности научных данных, кажутся самоочевидными, однако полезно было узнать, что же именно думают специалисты. Учёных попросили дать собственную оценку («очень важно», «важно», «относительно важно», «совсем неважно») каждому из следующих семи мотивов, или позиций:

1. Если исследование выполнено за счёт госбюджета, его результаты становятся общественной собственностью и должны соответственно сохраняться.

2. Это будет стимулировать продвижение науки (новые исследования могут строиться на базе существующего знания).

3. Может оказаться полезным для проверки в будущем.

4. Способствует повторному анализу существующих данных.

5. Стимулирует междисциплинарное сотрудничество.

6. Имеет потенциальное экономическое значение.

7. Это уникально.

По мнению учёных, возможность повторного анализа существующих данных – это наиболее важный стимул для организации системы их хранения: 91% респондентов оценили этот фактор как «важный» или «очень важный». Экономическая значимость рассматривается учёными самых разных дисциплин как наименее важная причина для обеспечения сохранности научных данных.

Наиболее серьёзным препятствием для налаживания технологии *data curation* учёные считают недостаток надёжного программно-аппаратного обеспечения: это отметили 80% респондентов. 58% признают необходимость формирования международной инфраструктуры для обеспечения сохранности и доступности научных данных; такая система поможет преодолеть отмеченное препятствие. 25% учёных уже разместили полученные ими данные в свободном доступе.

Основные препятствия для налаживания совместного использования научных данных учёные видят в нерешённости юридических вопросов и возможном неправильном использовании их собственных научных работ.

Чтобы иметь возможность планировать системы обеспечения сохранности научных данных, нужно знать: их состав, или виды (рис. 2), объём и длительность хранения силами самих учёных.

Рис. 2. Основные виды данных, используемых учёными

(допускалось несколько ответов)

Трудным для учёных оказался вопрос об объёме данных, которые они хранят, и о том, каким он окажется через пять лет. Около 10% респондентов не имеют представления об объёме существующих у них данных, а 17% – о том, сколько их будет через пять лет. В большинстве ответов указан примерный диапазон – от 1 Гб до 1 Тб. Полезно также отметить, что значительное число учёных (39%) к своим данным прикрепляют административную информацию (автор, дата создания, имя файла, происхождение и т.п.).

Не менее интересно местонахождение хранимых данных (рис. 3).

Рис. 3. Место, где учёные хранят свои научные данные (допускалось несколько ответов)

Учёные не слишком стремятся допустить совместное использование своих научных данных; только 25% респондентов объявляют свои работы открытыми любому пользователю; другие либо вообще не допускают к ним посторонних, либо допускают только ближайших коллег. Перечень препятствий к совместному использованию научных данных приведён на рис. 4. Суть в том, что учёные хотели бы сохранить контроль за собственными научными данными, и с этой точки зрения электронные архивы не имеют перспектив.

Мнения учёных о том, кто должен платить за обеспечение сохранности научных данных, представлены на рис. 5.

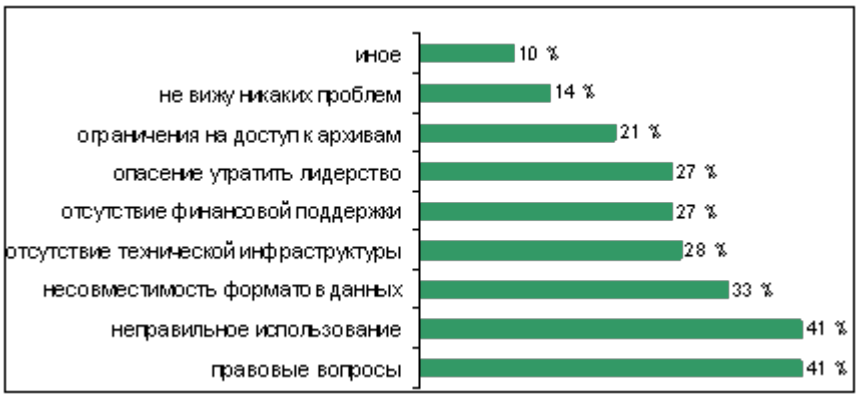

Рис. 4. Препятствия к совместному использованию данных

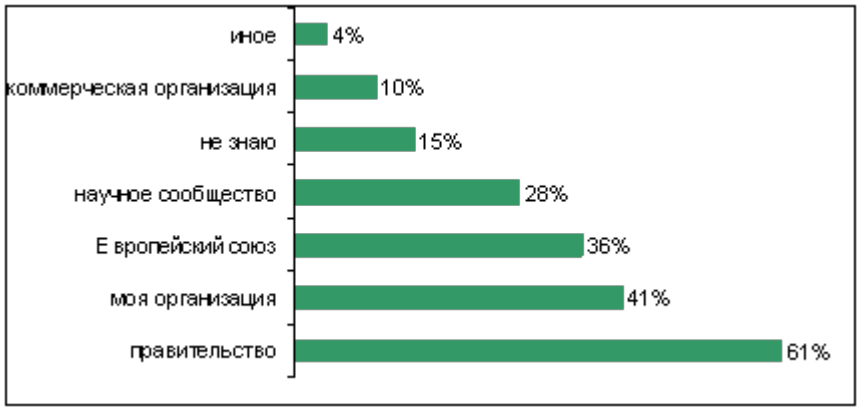

Рис. 5. Кто должен платить за обеспечение сохранности научных данных

Позиция специалистов по работе с данными. В этом докладе специалистами по работе с данными названы «профессионалы, имеющие ясно прописанные обязанности по обеспечению сохранности научных данных иработающие в научных библиотеках, архивах и других организациях» (рис. 6).

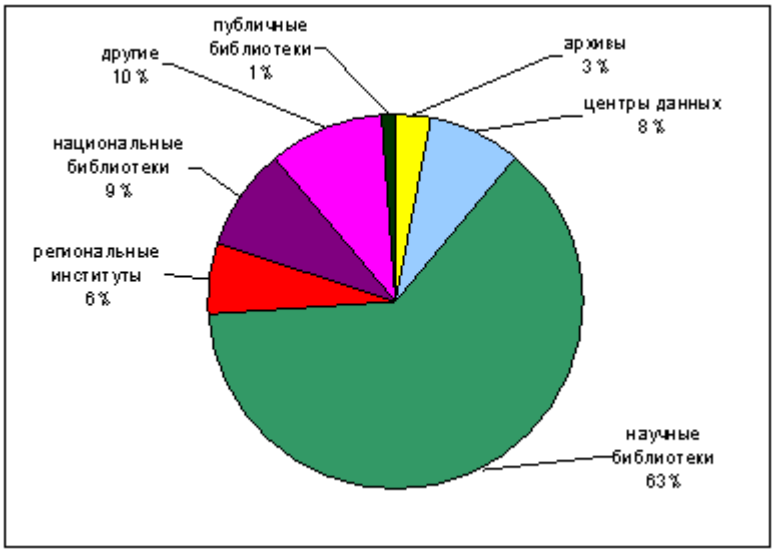

Рис. 6. Организации, работающие с научными данными

Виды электронных материалов, принятых на хранение соответствующими организациями, показаны на рис. 7, а типы и форматы таких материалов – на рис. 8.

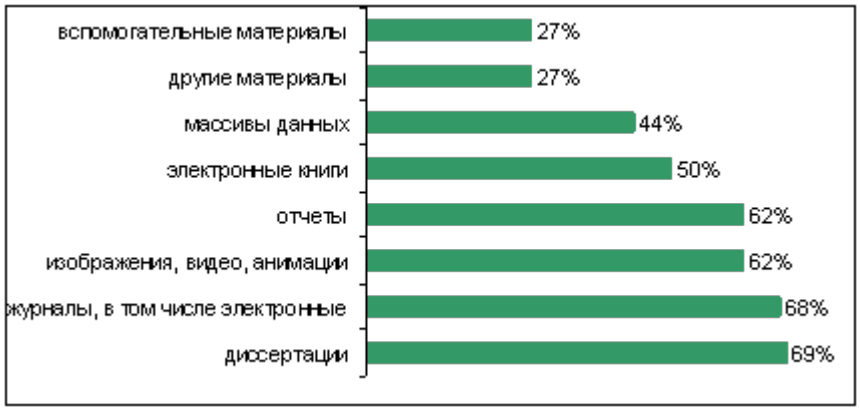

Рис. 7. Виды электронных материалов, принятых на хранение (допускалось несколько ответов)

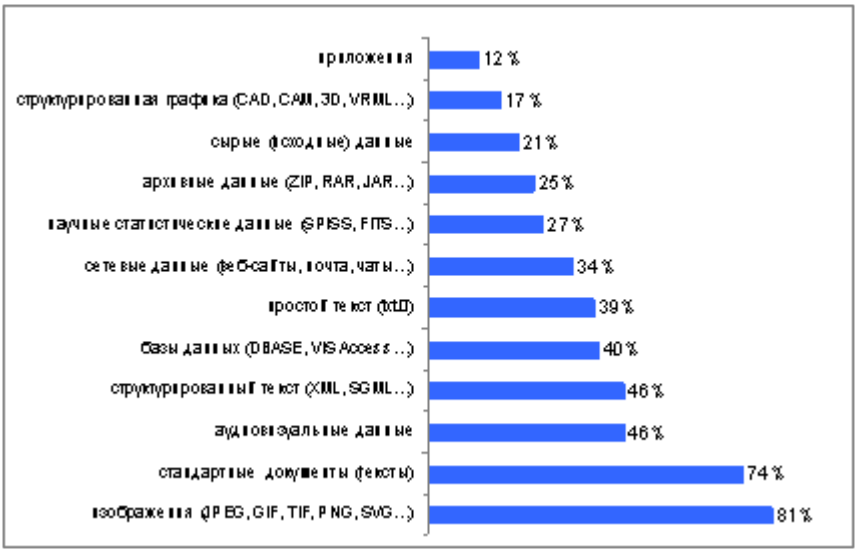

Рис. 8. Типы и форматы научных данных, принятых на хранение (допускалось несколько ответов)

98% опрошенных считают: если исследование профинансировано из госбюджета, то его результаты являются общественным достоянием и, следовательно, должны сохраняться.

Большинство (86%) специалистов по работе с данными наиболее серьёзными препятствиями для налаживания технологии *data curation* считают нехватку надёжного программно-аппаратного обеспечения, слабую поддержку от компьютерного сообщества; 60% признают необходимость формирования международной инфраструктуры для обеспечения сохранности и доступности научных данных – такая система поможет справиться с вышеназванными препятствиями. 59% не считают удовлетворительным нынешнее состояние инструментов и инфраструктуры для реализации поставленных целей. 71% полагает, что финансирование процессов сохранения научных данных будет проблемой на ближайшее будущее (до пяти лет).

Оценки объёмов данных, которые хранятся в настоящее время или будут храниться через пять лет, заметно различаются, однако наиболее часто назывались цифры в диапазоне от 1 терабайт до 1 петабайт.

Интересна позиция специалистов по поводу возможности связать данные, которые хранятся в их организации, с той публикацией, где изложены результаты обработки этих данных (рис. 9).

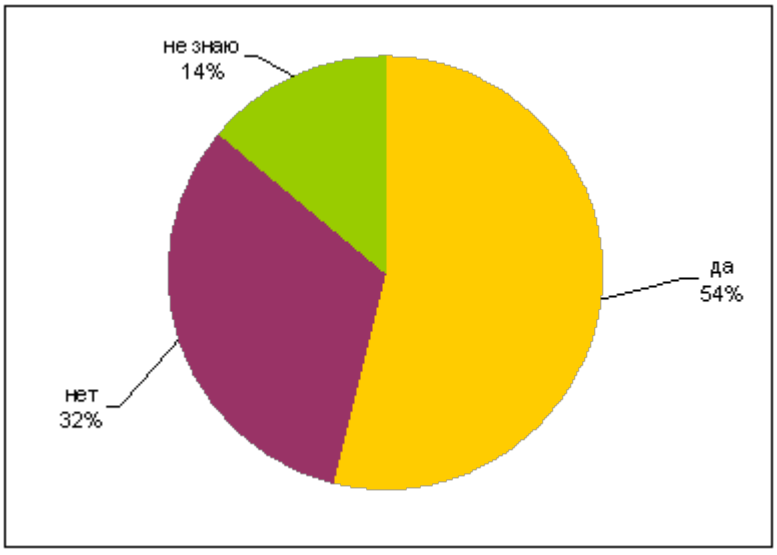

Рис. 9. Ответы на вопрос о возможности связать данные, которые хранятся в организации, с той статьей, в которой изложены результаты обработки этих данных

Только 46% опрошенных подтвердили возможность обмена данными между организациями, а не менее 50% не участвуют в обмене. При этом 89% считают, что такая технология необходима – налицо обычное расхождение слов и дела.

Позиция издателей. В мире публикуется около 25 400 наименований рецензируемых научных журналов (около 2 тыс. издающих организаций). При этом 5 крупнейших издателей выпускают 6 700 наименований, или 25% от общего количества. Подавляющее большинство – 88% крупных издателей и 95% средних – полагают, что хранить нужно научные статьи.

96% издателей считают наиболее важным стимулом для обеспечения сохранности то, что эта технология будет содействовать продвижению наук; до 80% называют основной угрозой сохранности – уход с рынка тех, кто заботится о научных данных в настоящее время.

Наименее существенной является экономическая значимость сохранения данных. Почти 75% считают полезным создание соответствующей международной инфраструктуры.

Пока что 69% издателей не имеют выработанных технологий; около 70% полагают, что авторы могут представлять научные данные как приложение к своим публикациям.

Наиболее вероятным сценарием развития издательского бизнеса респонденты считают гибридную модель, объединяющую подписную систему и открытый доступ к научным журналам.

Третий доклад – «Как "оседлать волну". Программа действий четырёх стран по проблеме научных данных» (A surfboard for riding the wave towards a four country action programme on research data; http://www.knowledge-xchange.info/Default.aspx?ID=4) – в ноябре 2011 г. подготовили Мориц Ван дер Граф (*Maurits van der Graaf*) из компании *Pleiade Managementand Consultancy* и Лео Вайерс (*Leo Waaijers*), консультант по системам открытого доступа, в соавторстве со специалистами из Голландии, Дании, Германии и Великобритании.

В Группу по обмену знанием (*Knowledge Exchange*), работавшую над докладом, вошли представители Электронной научной библиотеки Дании (*Denmark's Electronic Research Library*), Немецкого научного фонда (*German Research Foundation, DFG*), Объединённого комитета по информационным системам Великобритании (*Joint Information Systems Committee, JISC*), голландского фонда SURF.

Для учёных – основных производителей массивов научных данных – авторы доклада выделяют четыре мотивации к совместному использованию данных: 1. Повторное пользование и признание, 2. Этические соображения, 3. Требования финансирующих организаций, 4. Повышение доступности данных.

Обеспечение доступности больших массивов научных данных потребует совместных усилий ряда специалистов: во-первых, самих учёных, которым нужно приобрести базовые навыки работы с данными; во-вторых, людей новой, только ещё возникающей специализации – *специалист по данным*, которые должны отвечать за компьютерное обслуживание процессов хранения, доступность данных по соответствующей научной дисциплине, и в-третьих, представителей ещё одной новой специальности – *библиотекарей по научным данным*, т.е. тех, кто обеспечивает квалифицированное обслуживание научными данными, их сохранность и архивацию.

На основе анализа ситуации, сложившейся в четырёх странах, были сформулированы три долговременные стратегические цели: 1. Совместное использование данных должно стать частью университетской культуры; 2. Средства и инструменты хранения данных станут неотъемлемой частью профессиональной научной жизни; 3. Инфраструктура данных должна быть надёжной – как с оперативной точки зрения, так и с финансовой.

Насколько совместное использование научных данных распространено в мире? В Британском обзоре (*N. Beagrie, R. Beagrie, L. Rowlands; Research Data Preservationand Access: the views of researchers; Ariadne, 2009, nr. 60;* http://www.ariadne.ac.uk/issue60/beagrie-et-al/) отмечено: 22% респондентов, работающих в области естественных наук, 37% – в культурологии и в гуманитарных науках и 45% – в сфере общественных наук, пользуются научными данными совместно. В упомянутом выше обзоре *PARSE.Insight* сообщается, что 25% учёных готовы делиться данными со всеми.

Обзор университетов Дании (L. Waaijers, M. van der Graaf; Quality of research data, an operational approach; D-Lib magazine; Vol.17; nr.1/2; http://dx.doi.org/doi:10.1045/january2011-waaijers) содержит следующие сведения: 70% учёных, представляющих разные дисциплины, производят научные данные, из них 60% делятся ими с коллегами и 50% используют в работе «чужие» данные.

В Международном обзоре (C. Tenopir, S. Allard, K. Douglass, A.U.Aydinoglu, L. Wu, E. Read, M. Manoff, M. Frame; Data Sharing by Scientists: Practices and Perceptions; PLoS ONE ; Vol. 2011; http://dx.doi.org/doi:10.1371/journal.pone.00211) отмечено: только 36% учёных подтверждают, что найти их данные достаточно легко.

В этих условиях закономерно появление таких директивных документов, как «Принципы и руководства Организации экономического сотрудничества и развития для обеспечения доступа к научным данным, полученным в ходе исследований на государственные средства» (OECD Principles and Guidelines for Access to Research Data from Public Funding, 2007; http://www.oecd.org/dataoecd/9/61/38500813.pdf).

Европейский Союз ввёл понятие «Пятой свободы» – *свободное перемещение знаний*, в определение которого входят массивы научных данных.

В США некоторое время тому назад уже введены специальность «библиотекарь по данным» и тип библиотеки – «библиотека данных». Американские библиотеки закупают наборы данных у поставщиков (как правительственных, так и частных организаций) и обслуживают ими учёных.

В Великобритании *библиотека данных* создана в Эдинбургском университете.

Общая стоимость формирования инфраструктуры научных данных оценивается в 10-15% стоимости всей инфраструктуры науки. По данным комитета JISC (*Joint Information Systems Commetee*), институциональный репозитарий научных данных в 2,5–4 раза дороже, чем репозитарий для публикаций, поскольку в первом случае требуется большее количество персонала (обычно – до четырёх штатных единиц) и более дорогое оборудование.

Билефельдская конференция 2012 года

Доклады, представленные на Билефельдской конференции этого года (http://conference.ub.uni-bielefeld.de/programme/), развивают и конкретизируют положения обзорных докладов по рассматриваемой тематике.

В докладе Вильяма Мишенера (*William K. Michener*) из Университета штата Нью-Мексико (Альбукерке, США) «Новые парадигмы сырых данных для науки и вузов» (New Research Data Paradigms for Science and Academia) отмечены пять трендов, влияющих на смещение к новым парадигмам:

1. Сегодня научные данные рассматриваются как ценный продукт научной работы;

2. Библиотеки во всё большей степени работают с электронными документами, и постепенно наступает новая эра репозитариев для информации, знаний и данных;

3. Серьёзные проблемы ставит «большая наука», которая доминирует в исследованиях;

4. Широко ведутся исследования с интенсивным потоком данных;

5. Управление научными данными становится потребностью.

Научные исследования всё больше фокусируются на долговременных, широкомасштабных и сложных проблемах, для решения которых используются огромные массивы различных данных. *DataONE (Data Observation Network for Earth*) – это новая платформа киберинфраструктуры, разработанная для поддержки быстрого поиска данных и организации доступа к различным центрам научных данных.

В докладе Даана Бродера (*Daan Broeder*) из Института психолингвистики им. Макса Планка (Ниймеген, Нидерланды) «Строительство новой европейской инфраструктуры данных» (*EUDAT – Buildingan European Data Infrastructure*) подчёркнуто: представители европейского научного сообщества признают

необходимость обрабатывать постоянно возрастающие массивы научных данных, поступающих из различных источников – новых сенсоров, новых научных инструментов, которые возникают при компьютерном моделировании и оцифровке библиотечных фондов. Сегодня пока ещё нет действенной поддержки в работе с этими данными. Проект EUDAT задуман как панъевропейская сервисная инфраструктура по работе с научными данными, помогающая организовать обмен между репозитариями, центрами высокопроизводительных вычислений, каталогами метаданных и т.п.

Во многом схожая с ГПНТБ России Национальная немецкая библиотека по науке и технике (TIB/UB, Гановер) представила доклад, посвящённый проекту Международного консорциума для широкого использования научных данных (*DataCite – International Consortium for Data Citation*); с докладом выступил Ян Бразе (*Jan Brase*).

Сегодня глобальные научные исследования, как правило, ведутся коллективами учёных со всего мира, сопровождаются передачей потоков информации в масштабе всей планеты и поэтому нуждаются в глобальных международных стандартах. Однако на практике исследование выполняется локально, местными специалистами в рамках локальной инфраструктуры и при поддержке локальных финансирующих организаций.

Международный консорциум *DataCite*, основанный в 2009 г., объединяет 16 институтов из 12 стран. Его задача – совершенствование системы научного цитирования и повторного использования научных данных. С помощью постоянного идентификатора цифровых объектов DOI зарегистрировано более 1 млн массивов научных данных, поэтому можно: публиковать их в качестве независимого научного объекта, ссылаться на них, связывать их с журнальными статьями, отслеживать пользование ими.

Профессор Филипп Чимиано (*Philipp Cimiano*) из Билефельдского университета представил доклад «От связанных данных – к связанной науке» (*From Linked Datato Linked Science*). Парадигма «связанных данных» (*The Linked Data paradigm*) получила развитие в последние годы как отражение процессов публикации больших массивов данных.

Принципы «связанных данных» были сформулированы Тимом Бернерсом-Ли (*Tim Berners-Lee*, http://www.w3.org/DesignIssues/LinkedData.html) применительно к формированию Сети научных данных, которая смогла бы обеспечить доступность и повторное использование данных. Однако процесс создания такой сети идёт достаточно медленно.

Генеральная конференция ИФЛА в Хельсинки, 2012 г.

Три сообщения, прозвучавшие на организованной Секцией научно-технических библиотек сессии «Роль библиотек в работе с данными, обеспечение доступности и сохранности: международные перспективы» (*The role of libraries in data curation, access and preservation: an international perspective*), по договорённости с их авторами и с одобрения аппарата ИФЛА переведены и опубликованы далее в этом номере. Как вступление к этим докладам приведу лишь некоторые соображения.

Data curation, или хранение и обслуживание научными данными, получило широкую поддержку в мире, особенно после представления рассмотренного выше доклада «“Оседлать волну”. Каким образом Европа может получить пользу от нарастающей волны научных данных. (Финальный доклад Группы высококвалифицированных экспертов по научным данным)», выполненного по заказу Европейской комиссии в октябре 2010 г.

Заинтересованность Еврокомиссии нашла отражение в докладе Сюзан Рейли (*Susan Reilly*) – представителя Ассоциации европейских научных библиотек (LIBER) – «Роль библиотек в поддержке обмена научными данными»(*The role of libraries in supporting data exchange*).

При внимательном изучении *data curation* заметно, что движущей силой этой инициативы являются, скорее, библиотекари и специалисты по информации, а не сами учёные. Неравнодушные, думающие библиотекари изо всех сил стараются возродить роль и значимость библиотек.

Один из главных доводов в пользу этой новой технологии – возможность повторного использования так называемых сырых данных. Для реализации *data curation* Еврокомиссия и другие финансирующие организации готовы предпринять серьёзные усилия: переобучать библиотекарей, налаживать специальное

взаимодействие библиотекарей и учёных, разработать новую систему метаданных специально для работы с «сырыми» научными массивами и т.д. Однако дело движется неспешно. В своём докладе «Роль библиотек в обеспечении сохранности и обслуживании научными данными в Германии: результаты исследования» (*The role of libraries in curation and preservation of research data in Germany: findings of a survey*) Ахим Освальд (*Achim Osswald*) из Университета прикладных наук (Кёльн) и Стефан Стратман (*Stefan Strathmann*), сотрудник библиотеки Геттингенского университета, отмечают:

1. Немецкие библиотеки и библиотекари пока что находятся на предварительной стадии работы по сохранению и обслуживанию «сырыми» научными данными.

2. Сотрудничество библиотекарей и библиотек с учёными является существенным условием успеха.

3. Программы обучения студентов и аспирантов должны быть полностью перестроены, для того чтобы коренным образом изменить ситуацию.

4. До тех пор пока в библиотеку не придут специально обученные методам работы с научными данными профессионалы, необходимо организовать переподготовку действующего персонала.

Схожие доводы приведены и в совместном докладе «Работа с научными данными: мнения и позиции сотрудников вузовских библиотек» (Academic Librarians and Research Data Services: Preparation and Attitudes) Кэрол Тенопир (*Carol Tenopir*) – профессора Университета Теннеси (Ноксвилл), Роберта Сэндаски (*Robert J. Sandusky)* – доцента Университета штата Иллинойс (Чикаго), Сюзи Аллард (*Suzie Allard*) – доцента, заместителя директора Школы информационных наук Университета штата Теннеси (Ноксвилл), и Бена Берч (*Ben Birch*), научного сотрудника того же университета. (Перевод доклада – далее в этом номере.)

Профессиональный и методический интерес представляет доклад директора по информационным технологиям Немецкой национальной библиотеки Рейнхарда Альтенхонера (*Reinhard Altenhoner*) «Библиотеки как поставщики сервисов для хранения электронных научных данных: практические выводы из проекта Немецкого научного фонда DP4lib – Электронная сохранность в библиотеках» (Libraries as servicee-brokers for digital data curation: Practical insights from the DFG project DP4lib – Digital preservation for libraries). Этот материал удачно дополняет недавно опубликованную в нашем журнале статью о практике работы Немецкой национальной библиотеки по комплектованию обязательного экземпляра электронных документов (Науч. и техн. б-ки. – 2012. – № 7. – С. 62–72).

Комментарии по теме

Поскольку по образованию я – физик-экспериментатор по ядерной физике (с 20-летним опытом работы), то позволю себе к этой инициативе, т.е. к *data curation*, отнестись немного скептически. Дело в том, что массив научных данных сам по себе не имеет большого значения. Исключения составляют лишь заранее сформированные коллективные проекты – такие, как геном человека, поиск бозона Хиггса, проекты в астрономии, совместные исследования климата Земли. В этом случае заранее обговариваются методики, инструментарий, допустимый уровень погрешностей и т.п.

Если же не было договорённости *априори*, то трудно понять, зачем всё это нужно. Сами по себе данные – без описания конфигурации эксперимента и других многочисленных сопутствующих факторов – не представляют большого интереса, и зачастую проще сделать всё заново, чем восстанавливать «как это было». Надежды на активное участие учёных в повторном использовании научных данных также, на мой взгляд, довольно призрачны. Во-первых, основной движущей силой настоящей науки является любознательность, т.е. стремление человека узнать новое об окружающем нас мире, а не повторять чужие открытия (академик Лев Андреевич Арцимович в шутку говорил нам: «Наука – это удовлетворение собственного любопытства за государственный счёт»). Во-вторых, технологии исследований так быстро развиваются (во многом они связаны с продвижением информационно-коммуникационных технологий), что трудно себе представить особый интерес к старым данным.

Столь же удивительно стремление развивать систему метаданных специально для работы с массивами данных, т.е. с набором цифр. Полагаю, что достаточно использовать те метаданные, которые были сгенерированы для печатной статьи об этом эксперименте. Объём публикации, как правило, заранее

ограничен редакцией (например, не более 3 с. или 20 с.). На электронную версию статьи такие ограничения не распространяются, поэтому вполне возможно все наборы «сырых» исходных данных и вообще всё, относящееся к эксперименту или наблюдению, включить как естественное дополнение к основному тексту и (в первом приближении) считать созданные метаданные неотъемлемой частью печатной статьи, а также электронной версии со всеми её приложениями. И никакой огород заново городить не нужно – это пустая трата сил и средств.

Ещё одно наблюдение. Последние 3–4 года девизом передовых библиотек было: «От формирования фондов – к насущным потребностям пользователя». Опубликовано множество интереснейших статей об изучении этих потребностей, о том, сколь неэффективен и старомоден принцип создания коллекций «на всякий случай» – даже тщательно сформированные фонды используются в университетах на 30–40%. И вот – резкий поворот: люди готовы тратить массу усилий на сбор, обработку и хранение того, что может быть никогда и не понадобится. Это диалектика или метание из стороны в сторону? Предложения по сбору и вторичному использованию «сырых» данных (*raw data*), на мой взгляд, нуждаются в доработке и оценке. Если не установить жёстких критериев селекции, то поток поступающей на хранение информации (массивов «сырых» данных) приведёт к «информационному загрязнению» (*information pollution*). Очень важно убедиться в фактической востребованности «сырых» данных, набрать полноценную статистику повторного использования. Пока что шумиха вокруг этого направления очень напоминает мне ситуацию с «проблемой тысячелетия» Y2K – много слов, мало фактов.

Однако навязывать своё мнение всему библиотечному сообществу было бы неправильно, тем более, что наука всё-таки сильно изменилась, и объём численных данных колоссально вырос.

Именно поэтому редакция нашего журнала, получив разрешение, подготовила перевод трёх наиболее типичных выступлений, посвящённых работе с «сырыми» данными, на генеральной конференции ИФЛА в Хельсинки в 2012 г. – Сюзан Рейли, Кэрол Тенопир с соавторами и Рейнхарда Альтенхонера. Обращаем внимание наших читателей и на методические аспекты. Все три работы подготовлены специалистами высокого международного уровня и финансировались очень богатыми организациями. С. Рейли действовала в рамках 7-й Рамочной программы Европейского Союза, а команда К. Тенопир готовила обзор по заказу Национального научного фонда США. Должность Р. Альтенхонера – директор по ИТ Немецкой национальной библиотеки – говорит сама за себя. Так что у них есть чему поучиться.