Устойчивое обеспечение сохранности в различных контекстах [1]

2. Данные исследований

В то время как научный дискурс объединяет результаты научной деятельности, эти исследований включают в себя исходные данные для новых исследований, а также результаты первого порядка. В результате развития цифровых технологий исследования практически во всех областях человеческого знания трансформировались. Научный инструментарий и информационные технологии расширили возможности наблюдения, документирования, моделирования, и мы сегодня можем находить, анализировать, визуализировать и представлять какими-либо иными способами больше данных, чем это было возможно с традиционными носителями информации. Некоторые наблюдатели называют новые режимы *исследовательской работы информационно-интенсивными* [2]. Как результат возросла значимость таких факторов, как доступность, целостность, управляемость данных, существенно изменилась практика архивирования почти во всех областях исследований.

Скорость появления инноваций в информационно-интенсивных исследованиях огромна, поэтому методы управления этим контентом, с которыми приходится сталкиваться сегодня, завтра уже изменятся. Стратегии и практические методы должны быть достаточно гибкими, чтобы быстро адаптироваться к изменениям технологий, критериев отбора, способам использования данных. Мы сосредоточимся, главным образом, на естественных и общественных науках, но прежде отметим, что информационно-интенсивные исследования столь же быстро трансформируют гуманитарные науки, которые становятся все больше зависимыми от первичных цифровых ресурсов. Поэтому все, что мы здесь скажем, будет относиться также и к гуманитарным дисциплинам.

2.1. Ценность и отбор

Данные научных исследований весьма разнообразны – по типу и объему, использованию и долгосрочной ценности. Мы рассматриваем четыре вида исследовательской информации, обладающие характерными для них содержательными атрибутами.

Данные наблюдений телескопов, спутников, сенсорных сетей, а также геофизических исследований и другая информация исторического или единовременного характера (например, SDSS, астрономические данные Слоуновского цифрового обзора неба). К этой категории относятся также данные социальных исследований (например, демографические исследования,

[1] Продовження. Початок. див.: Шляхи розвитку української науки. – 2014. – № 3. – С. 70–74; Шляхи розвитку української науки. – 2014. – № 4. – С. 77–80.

[2] См., например, Beyond the Data Deluge // Gordon Bell, Tony Hey, and Alex Szalay, Science Magazine 323, March 2009, p. 1297–98.

проводимые ICPSR – Межуниверситетским консорциумом политических и социальных исследований при Мичиганском университете). Во многих случаях такие данные не могут быть получены повторно и поэтому должны быть сохранены.

Экспериментальные данные получают с использованием высокопроизводительного оборудования (например, ускорителей) путем клинических исследований, биомедицинского и фармацевтического тестирования, и других контролируемых экспериментов. Особенно важно обеспечить сохранность экспериментальных данных, в тех случаях, когда повторный сбор данных невозможен или неэтичен. В некоторых случаях такие данные касаются человека как объекта исследований и исчезающих видов животных.

Данные вычислений – результаты крупномасштабных проектов по математическому моделированию объектов. Несмотря на то, что данные эти можно воспроизвести путем повторного моделирования, существуют две причины, по которым их следует сохранять на среднесрочный период (три года и более). Во-первых, данные могут быть использованы как основы независимого и последовательного анализа, визуализации, «добычи знаний». Во-вторых, иногда в нужный момент может не оказаться компьютерного времени для дополнительных вычислений. Очень часто для самых крупномасштабных вычислений необходимы суперкомпьютеры, которые обслуживают всю страну, например, суперкомпьютеры национальных лабораторий Министерства энергетики США *(U. S. Department of Energy)* и центров Национального научного фонда США *(National Science Foundation, NSF).*

Справочные данные требуют интенсивного сопровождения и пользуются большим спросом у многочисленных научных сообществ. Такие данные создаются в самых различных целях – от картирования человеческих генов и описания белков до многолетнего сбора данных по экономическим и социальным вопросам. Международный банк данных белков *(Worldwide Protein Data Bank)* и Панельное исследование динамики доходов *(Panel Study of Income Dynamics)* – вот примеры такого рода данных [3].

Помимо этих данных иногда необходимо сохранять и вспомогательную информацию, например, о калибровке инструментов, экспериментальных параметрах, данные лабораторных журналов.

Наиболее крупные массивы исследовательских данных созданы и используются в науке, однако и в политике такие данные играют важную роль. В государственной политике используются данные по климату, сейсмологии, океанографии, клиническим и социальным исследованиям, исчезающим видам

[3] См. http://psidonline.isr.umich.edu.

животных и растений, заповедным территориям и археологическим раскопкам, вопросам безопасности, – все это выходит за рамки науки и представляет собой приоритетные вопросы государственной политики. Это вторичное использование данных исследований указывает на то, что в долгосрочной перспективе потребность в обеспечении доступа к ним также будет существовать.

Неопределенность будущей ценности. Профессиональные общества и другие признанные уполномоченные организации играют важнейшую роль в решении вопроса о том, что следует сохранять и в течение какого времени. Наиболее устойчивой стратегией, с которой нам пришлось столкнуться, оказалась такая: доверенная организация или управляющий совет (соответственно, *ARC* и *ICPSR*, поскольку речь идет о данных астрономических наблюдений и социологических данных) была уполномочена сообществом давать оценки относительно приоритетов отбора. Со временем во всех информационно-интенсивных отраслях будет достигнут консенсус относительно критериев отбора, а придание полномочий доверенной организации – эффективный способ достижения и обеспечения отчетности.

Для каждой из отраслей должно быть ясно, что новые данные, более полные массивы данных и новый инструментарий, – все это обесценивает данные, полученные в предшествующие периоды. Решение о том, что *не* следует сохранять, иногда не менее важно, чем решение о том, что должно быть сохранено. В некоторых случаях существует убедительная причина, заставляющая хранить информацию бессрочно, как, например, в гуманитарных дисциплинах, а также тогда, когда долгосрочное хранение продиктовано этическими соображениями. В других случаях – когда речь идет о вычислениях или наблюдениях – старые данные обесцениваются с появлением новой информации, например, данные последующих испытаний, показания более точного прибора.

Рекомендация № 1. Каждая отрасль, действуя через профессиональные общества или совещательные органы, должна определить приоритеты отбора данных, интенсивность их сопровождения, длительность хранения.

2.2. Стимулы к сохранению

Недостаточное стимулирование. Иногда создатели информации не имеют весомых стимулов к ее сохранению, даже несмотря на признание ее долгосрочной ценности и хорошо разработанные критерии отбора [4]. Практика показывает, что стимул уменьшается по мере понижения уровня принятия

[4] *Hedstrom M.* Incentives for Data Producers to Create «Archive-Ready». Data: Implications for Archives and Records Management. – Mode of access: http://files.archivists.org/conference/2008/researchforum/HedstromNiu-AbstractBio.pdf; *Amy. M. Pienta.* The LEADS Database at ICPSR: Identifying Important «At Risk» Social Science Data. – Mode of access: http://www.icpsr.umich.edu/files/DATAPASS/pdf/Pienta_et_al_2008.pdf.

решений – на самом низком уровне находится индивидуальный исследователь. В целом и в частности обеспечение сохранности результатов исследований, проводимых по грантам, – это работа, в которой никто не заинтересован. Время и деньги на обеспечение этой деятельности вычитаются из общего бюджета на исследование. Это существенно влияет на мотивацию, независимо от того, вменено ли обеспечение сохранности в обязанность или нет.

Обычно для стимулирования к сохранению информации рекомендуют именно включение этой деятельности в чьи-то обязанности, в частности финансирующие организации обязывают стипендиата сохранять результаты исследований. Чтобы эта обязанность выполнялась эффективно, необходимо четко распределить средства (например, выделить на эти цели часть гранта). Также необходимо иметь четкие критерии отбора – не может быть обязанности сохранять все подряд, – а также нужен партнер, с которым может сотрудничать основной исследователь. Каждой отрасли необходимы партнерские отношения со специалистами по данным или информатике, знающими именно эту отрасль, с тем, чтобы данные имели эффективное сопровождение и были должным образом подготовлены к депонированию в архив.

Рекомендация № 2. В необходимых случаях финансирующая сторона должна вменять обеспечение сохранности в обязанность получателя средств. Формулируя эту обязанность, необходимо указать критерии отбора, объем финансирования и назвать организации, ответственные за архивирование информации.

Финансирующей стороне следует способствовать наращиванию возможностей исследовательских институтов в этой области. Начать работу можно с университетских библиотек, которые могли бы эффективно распоряжаться сохраненными данными исследований. Также важно понять, какие компетенции и мощности требуются в действительности. Такие агентства, как *NSF, JISC*, Национальный институт здравоохранения (*National Institutes of Health, NIH)*, а также *Wellcome Trust* (медицинский благотворительный исследовательский центр в Великобритании), могут совместно с грантополучателями и исследователями в своей отрасли определить потребности в архивировании информации и принять решение о том, как может быть достигнут эффект масштаба при том или ином специализированном сценарии. Федеральные агентства также могут обеспечить первоначальное финансирование разработки и внедрения архивного оборудования как важнейшей части устойчивой исследовательской инфраструктуры.

Каждое финансирующее агентство должно искать способ указать на значение управления данными. Например, *NSF* и *JISC* могли бы сделать управление сохраненными данными базовым научным показателем и признать эту категорию данных фундаментальным научным ресурсом. Для этого

необходимо будет регулярно и в стандартизованной форме оценивать, сколько данных произведено, сохранено, видоизменено и повторно использовано.

Рекомендация № 3. Финансирующим агентствам следует признать «управление сохраненными данными» базовым научным показателем, включив его в стандартные формы отчетности.

2.3. Функции, ответственность, финансирование

Стимулирование не означает автоматически финансирования. Средства, необходимые для обеспечения сохранности, часто принимают форму наличной оплаты, единовременных грантов, добровольной бесплатной работы (как в случае *wwPDB*), и поэтому эта деятельность не слишком надежна. На сегодняшний день финансирование является относительно гарантированным лишь в тех случаях, когда принята модель подписки (например, модель, принятая Межуниверситетским консорциумом политических и социальных исследований, *ICPSR*) и решена проблема неоплачиваемого использования, хотя последствием такого выбора является ограничение доступа для некоторых групп потенциальных пользователей.

Обычно существует взаимосвязь между максимально возможным обеспечением доступа и максимальным финансированием. В случае с проектом Слоуновского цифрового обзора неба *(SDSS)* обеспечение максимального доступа является решающим для астрономии условием, поэтому естественно, что научное сообщество само предоставляет финансирование. Проблема неоплачиваемого доступа потенциально существует и для *SDSS*, однако в астрономическом сообществе развита культура совместного использования ресурсов и сосуществования любителей и профессионалов. Лишь допустимо малая часть общего финансирования *SDSS* идет на обеспечение сохранности данных. Очень важно, что сам проект предусматривает механизмы переоценки приоритетов в отношении сохраняемой информации, что зафиксировано в Меморандуме о взаимопонимании, подписанном с партнерской архивирующей организацией. Институты, которые коллективно поддерживают *SDSS*, с большой долей вероятности будут продолжать поддерживать его и с появлением более широкого круга пользователей, даже несмотря на проблему неоплачиваемого использования.

Во многих случаях, когда сопровождение и архивирование информации требует наиболее глубоких профессиональных знаний в этой области, централизованное оказание таких услуг оказывается наиболее эффективным. В любой системе управления затраты на персонал составляют наибольшую часть общих затрат. Эти расходы могут быть снижены за счет автоматизации, однако полностью избежать их нельзя. При централизованном оказании услуг, как и во всех других случаях, заключается соглашение между сообществом и архивом с указанием условий доступа и результатов. В каждом соглашении

следует оговорить необходимость периодической переоценки, видоизменения или уничтожения данных. Также в этих соглашениях следует предусмотреть механизм передачи ресурса в другой депозитарий.

Рекомендация № 4. По возможности при обеспечении сохранности необходимо стремиться к сокращению расходов на сопровождение и архивирование данных за счет эффекта масштаба.

Рекомендация № 5. В соглашениях с третьей стороной – архивирующей организацией – должны быть указаны процессы, результаты, срок хранения данных, основания для передачи данных в другие руки.

2.4. Будущее

Главная трудность при принятии решений состоит в том, чтобы найти соотношение между текущим использованием и созданием новых данных, с одной стороны, и поддержкой надежного управления данными, с другой. В классической схеме в центре всех принимаемых решений лежит противоречие между затратами времени и ресурсов сегодня и необходимостью инвестирования в создание возможностей для будущего. Единого способа разрешения этого противоречия не существует. Тем не менее, заинтересованным сторонам необходимо понимать, что все информационно-интенсивные исследования фундаментальным образом изменяют долгосрочное распределение ресурсов между конкурирующими направлениями с целью поддержки устойчивого развития науки.

Блок 2. Программа действий в сфере исследовательских данных
• **Рекомендация № 1.** Каждая отрасль, действуя через профессиональные общества или совещательные органы, должна определить приоритеты отбора данных, интенсивность их сопровождения, длительность хранения. • **Рекомендация № 2.** В необходимых случаях финансирующая сторона должна вменять обеспечение сохранности в обязанность получателя средств. Формулируя эту обязанность, необходимо указать критерии отбора, объем финансирования и назвать организации, ответственные за архивирование информации. • **Рекомендация № 3.** Финансирующим агентствам следует признать управление сохраненными данными базовым научным показателем, включив его в стандартные формы отчетности. • **Рекомендация № 4.** По возможности при обеспечении сохранности необходимо стремиться к сокращению расходов на сопровождение и архивирование данных за счет эффекта масштаба. • **Рекомендация № 5.** В соглашениях с третьей стороной – архивирующей организацией – должны быть указаны процессы, результаты, срок хранения данных, основания для передачи данных в другие руки.

Возможно, с течением времени инфраструктура информационного обеспечения исследовательской деятельности и образования в естественных и общественных науках приобретет большее сходство с инфраструктурой информационного обеспечения гуманитарных наук, имеющей более долгую историю, – в том смысле, что оба эти направления основываются на многократном использовании ресурсов, которые доступны нынешним исследователям благодаря рациональному управлению наследием предыдущих поколений. Когда производство знаний основывается на использовании исторических, многолетних или уникальных данных, потребность в управлении ими возрастает. Эффект масштаба достижим не только в сфере естественнонаучных и социальных знаний, но и в сотрудничестве с гуманитарными науками. Естественнонаучные сообщества получают преимущества, связанные с долгой традицией управления историческими, уникальными материалами, а гуманитарные сообщества – с профессиональными знаниями и большим опытом специалистов в области информации, созданной при помощи машин ***(Устойчивая экономика для цифровой планеты: обеспечение долговременного доступа к цифровой информации. Итоговый отчет Рабочей группы по устойчивому обеспечению долговременной сохранности и доступа к цифровой информации [перевод с англ.]. – М.: МЦБС, 2013. – С. 118–129).***