МЕЖДУНАРОДНЫЙ ФОРУМ ПО ИНФОРМАЦИИ

2012 **Том 37** **№ 3**

Выходит ежеквартально **Основан в 1975 г.**

СОДЕРЖАНИЕ

Грехем М. Дж. Искусство науки о данных 3

Борнман Л., Лейдесдорф Л. Влияние цитирования статей, опубликованных в шести продуктивных странах: сравнение по странам, основанное на данных InCites 10

Лозано Ж. А., Ларивьер В., Жингра И. Ослабление взаимосвязи между фактором влияния и цитированиями статей в цифровой век 16

Франнсен Т.Ф., Николаисен Й. Если вы не можете найти журнал, существует ли он? Доступность журналов по библиотековедению и информатике в Интернете 22

Муньер П. Программа Freemium как поддерживающая экономическая модель для электронного издательства открытого доступа в области гуманитарных и общественных наук 29

Москва **Всероссийский институт научной и технической информации РАН (ВИНИТИ)**

РЕДАКЦИОННАЯ КОЛЛЕГИЯ

Академик РАН **Ю.М. Арский** (Российская Федерация) — *главный редактор*, ВИНИТИ РАН, 125190, Москва, ул. Усиевича, 20. Телекс 411249

Проф. д-р. **Р.С. Гиляревский** (Российская Федерация) — *заместитель главного редактора*, ВИНИТИ РАН, 125190, Москва, ул. Усиевича, 20. Телекс 411249

С. Дж. Паркер (Канада) — *заместитель главного редактора*, IDCR, P.O. Box 8500, Ottawa, Ontario K1G 3H9, Canada

А. Джилкрайст (Великобритания) — CURA Consortium and GAVEL g.e.i.e, 38 Ship Street, Bringhton BN1 1AB, UK

М. Дрейк (США) — Технологический институт шт. Джорджия, Библиотечный и информационный центр, 704 Cherry Street, Atlanta, Georgia 30332-0900, USA

Н. Дюсулье (Франция) — 192 Chemin du Cannet, 06220 Vallauris, France

А. де Кемп (Германия) — Издательство "Springer-Verlag", Postfach 10 52 80, D-69042 Heidelberg, Germany

Д-р **Т. Кеннон** (Великобритания) — Отдел исследований и разработок Британской библиотеки, 2 Sheraton Street, London W1V 4BH, UK

М. Миддлтон (Австралия) — Школа информационных систем, QUT Gardens Point Campus, 2 George Street, Brisbane, 4000 QLD., Australia

Т. Молвиг (Норвегия) — Национальное управление по научной информации, вузовским и специальным библиотекам, P.O. Box 2439 Solli, N-0201, Oslo, Norway

Х. Ринкон Феррейра (Бразилия) — Бразильский институт информации по науке и технике (IBICT), SAS— Quadra 5, Lote 06, Bloco H, 700-70-000 Brasilia D.F., Brazil

С. Феррейро (Чили) — Чилийский университет, Системы информационных и библиотечных служб, Casilla de Correo 10D, Santiago, Chile

Проф. **Ю. Фудзивара** (Япония) — Университет Цукуба, Институт электроники и информатики, Tsukuba-shu,Ibaraki, 305 Japan

Д-р **М. Хименес** (Испания) — Испанское общество по научной документации и информации, Fuencarral, 123-6º dcha., 28010, Madrid, Spain

Д-р **А.И. Чёрный** (Российская Федерация) — ВИНИТИ РАН, 125190, Москва, ул. Усиевича, 20. Телекс 411249

© Всероссийский институт научной и технической информации РАН, 2012

Искусство науки о данных*

Мэтью Дж. ГРЕХЕМ
(Matthew J. GRAHAM)

Калифорнийский технологический институт, г. Пасадена, шт. Калифорния, США

В новой среде, интенсивно использующей данные науки XXI в., нам необходимо развивать новые навыки. Они могут выражаться терминами систематизированного подхода, формируемого на основе средневекового образования – тривиума (логика, грамматика и риторика) и квадривиума (арифметика, геометрия, музыка и астрономия). Однако данные являются скорее новым основным принципом, чем фокусированием на их числе. Нам необходимо понять, каким правилам они подчиняются, как они выражаются и передаются и какова их связь с физическим пространством и временем. Это понимание будет рассматриваться в терминах технологий и процессов, которых оно потребует. Утверждается, что, по крайней мере, важна оценка всех этих аспектов, чтобы дать нам возможность извлечь научную информацию и знание из массивов данных, которые угрожают поглотить и сокрушить нас.

ВВЕДЕНИЕ

Обучение в крупных средневековых университетах фокусировалось на семи свободных науках: *тривиум* - логика, грамматика и риторика, и *квадривиум* - арифметика, геометрия, музыка и астрономия. Подготовка и компетентность в этих предметах считались достаточными, чтобы сформировать у индивидуума необходимые интеллектуальные способности для продвижения в карьере или дальнейшего приобретения знаний в юриспруденции, теологии или натурфилософии. Современные натурфилософы обучаются искусствам эмпирической, теоретической и вычислительной научной методологии как подготовке для их профессиональной карьеры. Тем не менее, авангард революции данных обрушивает на нас сегодня высокоразмерные, многообъемные, обогащенные совокупности данных, которые становятся все более общим аспектом нашего повседневного рабочего места, а мы недостаточно подготовлены.

Чтобы принять этот вызов, появляется четвертая парадигма [1]: так называемая наука, интенсивно использующая данные, или *x*-информатика (где *x* является наукой выбора, такой как *био*информатика, *гео*информатика или *астро*информатика), которая будет поддерживать и продвигать научное открытие в XXI веке. Это не просто прирастающее развитие того, что было раньше, а нечто совершенно новое, и мы все еще пытаемся понять не только, какова его форма и где проходят границы, но более фундаментально, каковы его основные правила. Кроме того, в то же время это не было незнакомо ученому XIII века.

Сутью средневековой программы была систематизация знания – каким правилам оно подчиняется, как выражается и как оно передается, – и особенно числовое знание и связь числа с физическим пространством и временем. Арифметика, например, была наукой чистого числа, в то время как музыка была наукой числа относительно времени [2]. В этой статье мы стремимся показать, как новому искусству науки о данных может также быть придана форма систематизации *данных* и ее связи с пространством и временем, особенно в отношении их технологических аспектов. Это касается множества наук, но нашей основной темой будет астрономия.

ЛОГИКА ДАННЫХ

Как алхимики думали о ртути, что она *первичная материя* (первое вещество), из которой сформировались все металлы, так ученые полагают, что данные достаточны для полного понимания. Тем не менее, данные так же текучи и эфемерны, как их природный аналог. Большие затраты и усилия предприняты эмпириками, чтобы измерить их, компьютерщиками – чтобы имитировать их, и теоретиками – чтобы формулировать их, но даже в этом случае, понимаем ли мы реально, с чем работаем? Даже само слово открыто для размышлений [3]. Законное использование и особенно многоцелевое использование данных требует контекста: не только первичной обработки их сырых числовых или символьных значений, но и соответствующего внимания к их происхождению, систематике, тенденциям и границам.

Хогг и Ланг [4, 5] полагают, что большая часть астрономии проводилась через каталоги, второстепенный продукт данных, произведенный из сырых данных, но упускающий необходимое знание о данных – как оно анализировалось, ошибочно оценивалось и т.д., – чтобы поддерживать любое сложное статистическое умозаключение, такое как решение несвязанных вопросов в Слоуновском цифровом обзоре неба. Все значения, выходящие за рамки сырых данных, являются метаданными и нуждаются в достаточном описании, предпочти-

* Перевод Graham M.J. The art of data science— 2011.— http://arxiv.org/abs/1106.3305

тельно в терминах (бейесовой) последующей вероятностной модели так, чтобы произвольные вопросы (придуманные как гипотезы) могли быть заданы от них с максимальным использованием доступной информации. Приведенная в их экстремальных значениях окончательная модель должна состоять из всего неба через длину волны и времени, от которой любое астрономическое изображение, даже взятое в любое время с помощью любого оборудования в любой конфигурации, могло родиться и так, чтобы аномалии в любых данных можно легко установить.

Семантика обеспечивает альтернативный, но дополнительный подход, формируя знание о данных, скорее в терминах программных структур, чем функций правдоподобия. Семантические конструкции, такие как онтологии, позволяют знанию предметной области выражаться в машиночитаемом формате в терминах классов, свойств (связей) и операций и данных как их примеров. Логический вывод по классам и примерам позволяет определять противоречия как в данных, так и в их описаниях. Семантика также способствует многоцелевому использованию данных: различные описания/интерпретации могут эффективно сочетаться/согласовываться *машинами* для создания множеств данных более широкого масштаба и применимости, чем первоначально планировалось. Например, изучение распределений спектральной энергии, использующих множества многоволновых данных, образованных из наблюдений астрономических объектов путем сочетания (гетерогенного) единичного фильтра/полосы пропускания, требует соответствующей обработки значений потока в каждом множестве данных, информация которого должна быть закодирована в своих онтологиях или эквивалентных структурах.

Нам не следует просто слепо работать с данными, особенно потому, что это становится все более сложным. Исследования могут обходить полные и соответствующие виды обработки, но понимание правил, которым данные подчиняются, и манипулирование этой логикой через умозаключение, будь оно статистическим или логическим, необходимы для обоснованного и воспроизводимого открытия. Разработка таких систем и гарантия того, что они являются выполнимыми перед лицом предстоящих ожиданий данных, являются реальным вызовом, однако не тем, который можно легко истолковать, а тем, который может удовлетворяться междисциплинарным контактом через вертикальные шахты отдельных наук.

ГРАММАТИКА ДАННЫХ

Для средневекового разума разгадка тайн мира лежит в декодировании символических языков, которые применила природа, чтобы спрятать свои секреты. Все было предписано значению, будь это через число, цвет, геометрию или некоторый более тонкий аспект или свойство. Разумный человек мог прочитать скрытые сообщения (образцы в саду монастыря), тогда как глупый видел только формы (цветы), ничего не понимая больше их значений. Символизация данных является гораздо менее прозрачной: сложные объекты превращаются в последовательности битов для устойчивости и передачи, но существует еще разнообразие представлений (форматы последовательного упорядочения данных), каждое с особым значением и целью.

На своем основном уровне данные включают числа или символы, как правило, хранящиеся в цифровом (бинарном) представлении. Пока каждый фрагмент данных может только рассматриваться как аморфная часть битов, польза этого подхода в самом деле ограничена крупными объектами данных (блобами), такими как текущие мультимедиа, которые образуют растущую фракцию сетевого трафика. Данные являются более манипулируемыми, если они структурированы некоторым образом и описание этой структуры доступно. Они обладают даже большим преимуществом, если эта структура не зависит от какого-либо специального аппаратного оборудования или программного обеспечения и машинной обработки. Существует также различие между форматами, используемыми для необработанных данных, являющихся в основном бинарными, и метаданными и производными данными, такими как каталоги, которые более структурированы и преимущественно являются текстовыми.

Форматы необработанных бинарных данных стремятся быть предметно-специфическими, хотя наблюдается некоторое использование гибкой системы передачи изображений (FITS) за рамками астрономии. Вместе с другими форматами, такими как HDF 5 (формат иерархических данных 5), описания таких бинарных структур и их метаданных (часто в сочетании) являются особыми. Формат описания канала (CDF) и формат описания сетевого канала (netCDF) используют более широкое понятие, определяя общую модель данных для множества научных данных, которая имеет свой собственный ассоциативный прикладной программный интерфейс (API). Он оперирует данными для чтения, координатными системами, в которых отображаются данные, и особыми типами данных и отделяет данные о пользователе полностью от физических деталей своего запоминающего устройства.

Самыми известными представлениями текстовых данных являются расширенный язык разметки гипертекста (XML) и формат обмена данными (JSON), и существуют системы для описания структур этих данных, например XML Schema и JSON Schema. Тем не менее, их часто критикуют за то, что они являются неэффективными форматами, особенно там, где скорость обработки является фактором, поскольку она происходит на познаковой основе. Проблемы пропускной способности и хранения могут быть также важными, и бинарные версии совсем не лучше. Протокол Buffers поисковой системы Google следует подобным абстрактным путем к CDF/сетевому CDF и разработан как более быстрая альтернатива XML. Структуры данных определяются в терминах общей модели данных с прикладным программным интерфейсом (API) для доступа и манипуляции. Реальный формат базовых данных нематериален – значение по умолчанию является бинарным, но могут также использоваться текстовые форматы, – библиотеки обеспечивают необходимые интерфейсы для него. Инфраструктура удаленного вызова процедуры и последовательного упорядочения (Apache Avro) следует подобному подходу, применяя JSON для определения его структур данных, но только используя компактный формат бинарных данных.

Когда речь заходит о передаче данных, реальная физическая транспортировка – так называемый метод «на своих двоих» – остается одним из наиболее эффективных и надежных средств, жертвуя латентностью ради высокой пропускной способности, и применяемым во многих крупных астрономических проектах, а также коммерческими поставщиками услуг. Однако каждый пример остается заказным решением, игнорирующим стандартизацию, с точными деталями, известными только отправителю и получателю. При

желании передавать данные потенциальным миллионам пользователей в любом месте в любое время требуются альтернативные решения.

Несмотря на то, что мы живем во время наибольшей взаимной связанности в истории человечества, существующая инфраструктура недостаточна для наших потребностей: у нас официально закончились адреса Интернет-протокола IPv4, а Интернет-каналы напрягаются под давлением текущих метаданных. Усилиями следующих поколений, таких как Интернет2, разрабатываются усовершенствованные требуемые возможности, например, создание по запросу и планирование высокой пропускной способности, высокоэффективных каналов данных, но использование современной установки может быть также оптимизировано.

Стандартные технологии передачи данных полагаются на один поток/канал между отправителем/провайдером и получателем для переноса данных, который обычно не использует в полном объеме доступную пропускную способность. Ввод данных и отправка их через множественные потоки получателю находят гораздо большее применение для повышения пропускной способности каналов, например расширение стандартного протокола передачи файлов (GridFTP) работает таким образом. Эти потоки могут исходить либо от многочисленных провайдеров, каждый со своей собственной (частичной) копией данных (только требуемая часть данных должна быть доступна), или от одного провайдера, обслуживающего параллельные потоки. Части данных запрашиваются у провайдеров на основе их рекламной доступности, и если получатель имеет часть данных, он может стать также провайдером для них – это является основой многих транспортных систем передачи данных от одного коллеги другому.

Потоки данных обычно используют пакеты протокола управления передачей (TCP) для своей транспортировки, но это может иметь недостаточную эффективность на длинных расстояниях, особенно когда пропускная способность высока или когда в процесс вовлечены множественные конкурентные потоки с различными скоростями передачи данных. Единый стандарт передачи данных (UDT) использует пакеты пользовательского протокола данных (UDP) вместо того, чтобы достичь более быстрого переноса, чем может TCP, но с равной ему надежностью. Другие решения включают тонкую настройку TCP особенно для сетей высокой эффективности или измененного протокола TCP.

Не все форматы данных кодируют свою информацию настолько достаточно, насколько это достижимо, и часто можно сократить размер объекта данных для передачи (или хранения) путем его сжатия. Значительные улучшения могут быть получены, в частности, для текстовых данных с программами общей компрессии, такой как gzip и bzip2. Для астрономических бинарных данных – изображений и таблиц – компрессия мозаики FITS [6] предлагает лучшую эффективность, чем gzip и bzip2, и также сохраняет рубрики FITS (описания структур) несжатыми для более быстрого доступа. В действительности, что касается соответствующей библиотеки (CFITSIO), сжатые данные должны быть ошибочным методом для операции с декомпрессией, никогда не являющейся необходимой.

Что касается более крупных объемов данных, то хранение и пропускная способность становятся наградой – открытая архивная информационная система (OAIS) является высокоуровневой и высокоценной моделью для полного архивного цикла, полезной для создания дискуссий и критики относительно планирования управления данными. Значение данных, тем не менее, лежит в их наследуемой структуре и его независимость от действительной расстановки битов не отличается от абстрагирования значения объекта в мире, которым мы окружены.

РИТОРИКА ДАННЫХ

В средние века студенты механически обучались навыкам писать буквы и проповеди, используя риторические учения классической античности. Предполагалось, что структура языка соответствовала его состоянию и пониманию и, следовательно, манера и стиль хорошего и правильного общения были важны, используя подходящий тон и лингвистические конструкции для определенной предметной области (понимание, которое полезно для научного метода). Данные являются языком научной диалектики и высоко политизируются с набором ловких приемов и механизмов, чтобы привести аудиторию к определенному выводу.

Правдивость интерпретации является как функцией того, как она достигнута, так и вопросом веры в данные, на которых она основана. Степень такой веры, тем не менее, кажется, должна быть обратно пропорциональна тому, насколько легко доступны данные. Хотя астрономические данные являются коммерчески ничего не стоящими (описано Джимом Греем как проблема ценой в ноль млрд долл.), большая их часть до сих пор хранится в защищенных вертикальных башнях, доступных длительное время только для группы избранных лиц. Попытки создать открытую культуру данных рассматриваются как мятежные: единственной современной попыткой сделать эти данные изначально публично доступными является исследование [7].

Этот уровень контроля сохраняется, даже когда результаты подвергаются рецензированию и появляются в публичной сфере. Это может быть сизифов труд начать поддержку данных для воспроизведения или строительства на их основе частных интерпретаций. В науках о жизни, как правило, необходимо, чтобы были доступными без предварительных условий все данные, когда соответствующая статья публикуется, предпочтительно в публичном, одобренном сообществом хранилище данных. Астрономия уже обладает культурой центров данных, но они стремятся быть слишком привязанными к особым большим миссиям или длинноволновым режимам – определенно нет современного хранилища, в котором случайные данные могли бы архивироваться или даже постоянно регистрироваться.

Кристаллизованному прогрессу традиционного астрономического издательства противопоставляются интерактивные библиографические службы, такие как архив *arXiv* и система астрофизических данных (ADS), которые обеспечивают точки доступа к соответствующим данным, где возможно. Еще более современным направлением является предварительное обсуждение данных в блогах и других социальных сетевых форумах. Хотя большая часть этого является явно предназначенной скорее для общественности, чем для серьезного научного обсуждения, она отражает растущее разочарование существующей системой рецензирования, в частности, в нашем все более связанном и открытом обществе, и при нашем интересе к альтернативным способам, таким как открытое рецензирование и открытое комментирование коллегами.

Прогрессирующее появление междисциплинарных областей также является вызовом, так как данные часто вырываются из своего первоначального контекста и вторично используются в совершенно новой (и, возможно, не совсем подходящей) среде. Так называемая модель сбора и перемешивания (pick-and-mix model) допускает гораздо большую свободу для настоящих (несомненно) подтвержденных выводов либо намеренно, либо, более обычно, случайно в областях, где недостаточно развита экспертиза. Как упоминалось ранее, тем не менее, формальное использование семантики может прийти некоторым образом к предотвращению этого.

Тысячелетиями данные были ценным товаром, расположенным в избранных местах и сохраненным любой ценой. Необходимость открытого подхода в новую эру противостоит существующему контролю и структурам доступа, но гораздо больше гармонирует с намеренной чистотой и преданностью научному методу. Данные должны быть свободны для всех.

АРИФМЕТИКА ДАННЫХ

От счётов (абака) до алгоритма арифметика меньше касалась природы расчетов (вычислений), чем понимания природы числа, его свойств и уникальности числовых рядов, полученных путем определенных постоянных отношений. Это было гораздо более качественное, чем количественное изучение, мотивированное желанием предугадать присутствие незримой руки в природе, выраженной в красоте платоновского перфекционизма. Пока мы не ищем высшего смысла в данных, изучение их природы и свойств является все еще разъясняющим опытом.

Польза (или значение) данных лежит в их способности передавать информацию (хотя данные одного лица могут быть помехой другому лицу). Это высоко переменная величина, зависящая от размера и потенциального влияния содержания данных, т.е., насколько поддерживающими или вызывающими они являются для современной парадигмы, а также ее своевременности. Относительная польза отдельных фрагментов данных может ранжироваться, порождая общую тенденцию – логистику: начальные данные в этой области являются приблизительно экспоненциальными в полезности, например наблюдения 10 Ia-типов сверхновых звезд в спектре смещения красного $0{,}16 \le z \le 0{,}62$ предполагают ускоряющийся универсум [8]. Кроме того, так как всё больше данных становится доступным, происходит насыщение и их полезность снижается. Например, успешные наблюдения, поддерживающие результаты сверхновых звезд Ia-типа; и по достижении зрелости данные обладают нулевой полезностью – исследования регулярно показывают логистическое поведение. Метанаправление может стать последовательностью логистических поведений или чем-то приближающимся, что является в своем роде логистическим, зависящим от того, насколько новые парадигмы переопределяют пользу старых данных.

Беспрецедентный прогресс наравне с этими логистическими направлениями порождается двумя факторами. Во-первых, будущее скорее характеризуется массивными параллельными потоками (малых) событий данных, чем крупными монолитными пластами данных. Синергетическая природа данных (выраженная в законе Шандора Салаи – Szalay's law, что полезность N сравнимых наборов данных равна N^2) означает, что эти потоки четко приводят к потенциально быстрому прогрессу по логистической кривой при условии, что они являются соединяемыми. Парадоксально, но наступление эры, интенсивно использующей данные, отмечается точкой изгиба в росте полезности для отдельных наборов данных.

Во-вторых, наблюдается увеличивающаяся скорость приобретения данных, приводимая в движение экспоненциальным ростом скоростей в технологии (в частности, закон Мура относительно плотности интегральных схем в транзисторе). Некоторые люди верят, что эти скорости не могут увеличиваться неопределенно долго: на некотором этапе либо относительное удвоенное время работы различных технологий будет становиться несовместимым – самое медленное определяется пределом разрушения, – либо некоторые из них будут сталкиваться с резкой границей определённых физических законов, или экономика продолжающегося роста перестанет быть привлекательной или выгодной. Другие считают, что новые технологии будут появляться, чтобы, по крайней мере, сохранить экспоненциальный рост в эквивалентных долях, если не ускорить его.

Мощные принципы становятся все более важным аспектом. Уже в 2004 г. тактовые частоты микропроцессора передавали нерезонансные линии из-за пределов рассеяния мощности, хотя растущее число ядер на чип поддерживало процент роста вычислительной эффективности. Системы *Exascale* (настольный пета-флоп/встроенный терафлоп) предсказывали потребность в мощности ~ 100 MW [9], но даже процессоры на уровне продукта формируются в направлении к границам мощности. Одной смягчающей стратегией является применение графического процессора с большей общей целью вычисления, насколько это возможно [10], – они предлагают улучшенную флоп/ватт-эффективность, чем центральные процессоры. Однако они должны поддерживаться центральным процессором для запуска операционной системы и управления устройством графического процессора. Использование центрального процессора низкой мощности, который будет тратить большую часть своего времени на холостой ход, является важным краткосрочным решением, но неизбежно следует рассмотреть транссиликоновые технологии – они требуют меньше энергии, но ценой более медленной тактовой частоты.

Если универсум можно фундаментально сократить до набора уравнений, тогда существует определенное количество информации для извлечения из данных. Степень, к которой мы можем приблизить этот лимит, определяется технологией и энергией, доступными для нас в достижении этой цели, хотя, в конечном счете, закон сокращающихся доходов может пока толковать это как недостижимое. Однако, если мир непознаваем, тогда сбор данных и получение информации из него являются нескончаемой деятельностью.

ГЕОМЕТРИЯ ДАННЫХ

Крупные соборы средневековой Европы использовались как священные отображения творения, отражающие цель и структуру универсума через законы и формы геометрии, преобразованные главным каменщиком в имитацию работы его божественного мастера. Справедливо и то, что, огромные центры данных завтрашнего дня будут отражать стремления ведущих ученых и технологов через законы и формы новой геометрии облегчить изучение развития и структуры универсума, архитектурного порядка огромных массивов данных.

Физические средства священной геометрии хорошо понятны, будь то камень Каина и мрамор Пурбека или большие энергии. Системы хранения петабайтов могут конструироваться из компонентов продукта размером терабайт стоимостью приблизительно 50 тыс. долл. за буфер периферийного устройства на момент написания статьи, хотя должны быть приняты соответствующие предупреждения в целях защиты от высоких процентов неудач и последующей потери данных, которые ассоциируются с «дешевыми» товарными дисками. Искусство и умение в таком случае лежат в разделении данных на те, которые достаточны и эффективны для пользователей, и те, которые им противоположны.

Стандартная архитектура для высокой пропускной способности данных, которая предназначена быть преимущественно читаемой и редко перезаписанной или прилагаемой (например, для обработки данных), должна разделиться на фиксированного размера фрагменты данных (обычно 64 мегабайта). Затем можно будет распространить многочисленные копии (обычно три – две на одной стойке и одну на другой) каждого фрагмента данных вдоль дискового кластера (см., например, *Google FS* [11] или его эквивалент открытого источника, *HDFS*). Это обеспечивает надежность против потенциальных недостатков соответствующей аппаратуры и может быть точно отрегулировано (большее число копий) для особых данных, где ожидаются более высокие требования или защита. Центральный/главный узел поддерживает список того, где какой фрагмент и любые ожидаемые данные находятся, а также список всех операций, включающих данные. Тем не менее, это представляет очевидную единственную точку неудачи и может ограничить расширяемость данных (распределение главного узла является возможным решением).

Такие системы оптимизированы для очень крупных наборов данных с небольшим числом включенных частей. Когда имеются большие количества маленьких файлов в наборе данных, доминирующий процесс во время выполнения рабочего цикла вычисления на этом наборе данных располагает релевантными фрагментами, т.е. трафиком к главному узлу [12]. Система *HDFS* смягчает это путем определения структуры особых данных для таких ситуаций – файл последовательностей, который является, по сути, контейнером меньших файлов, связанным с индексом, – значительно сокращая число файлов на диске, необходимых для обработки. Дальнейшие улучшения могут быть достигнуты структурированием файлов последовательностей в соответствии с некоторым предписанием, т.е. пространственным или временным расположением графических файлов, а не случайным группированием файлов в них.

Альтернативный сценарий данных включает низколатентный случайный доступ (высокая доступность) к данным, например, поиск небольшого изображения или очень большого количества разноразмерных файлов с многочисленными конкурентными записями, например системными журналами. Для этих случаев больше подходят системы, основанные на распределенных многомерных картах, таких как Google's BigTable [13] или эквивалент открытого источника Hadoop, нереляционная распределенная база данных HBase (обе построены на методах архивирования GFS и HDFS, соответственно) или на более общих распределенных данных и архитектурах метаданных, таких как OpenStack или Swift iRODS.

Все эти физические архитектуры не имеют знаний о структуре данных, которые они используют. Тем не менее, существует подкласс, касающийся специального типа данных, который обычно можно положить в (реляционную) базу данных (RDBMS). Реляционные базы данных плохо функционируют за пределами объемов ~ 100 терабайт [14], но имеется четкая потребность в эквивалентных системах с целью поддержать петамасштабные каталоги, и т.д. Карта *BigTable* и ее варианты принадлежат к суперклассу систем, известных как ряд проектов, направленных на реализацию моделей баз данных (NoSQL), которые обеспечивают распределенное хранение для структурированных данных и могут использоваться как шкалированные эквиваленты к базам данных для многих типов данных. Тем не менее, лучшее соответствие для научных данных достигается системой управления базами данных (SciDB), которая является ориентированной на столбец (а не на строку, как RDBMS) и использующей матрицы в качестве объектов первого класса, а не таблицы, и до сих пор является автоматической классификацией и интерпретацией данных (как RDBMS, но не похожей на большинство решений NoSQL).

Сложные геометрии, которые мы применяем в наших центрах данных с дублирующими иерархическими моделями, не отличаются от тех, что использовались каменщиками десятки веков назад в их собственных возвышающихся сооружениях. И те, и другие должны показать наше знание цели и структуры самой вселенной, отраженные в творчестве людей.

МУЗЫКА ДАННЫХ

Древние люди верили, что небеса наполнены гармонией сфер, величественной фугой, созданной движениями небесных тел. Средневековый курс обучения оформлял это, наравне с внутренней фугой человеческого тела и слышимыми фугами, которые мы создаем, в понятие *музыки*, которая изучала движение пропорций во времени в соответствии с хорошо установленными образцами и правилами. Движение данных сквозь время как результат вычислений, производимых на них, представляет подобную фугу, и в случае больших массивов данных имеется ряд определяемых моделей.

Такой широко распространенной вычислительной моделью сегодня является так называемая смущающая параллельная задача, описывающая вычисление, для которого требуется мало или вообще не требуется усилий, чтобы выделить его в ряд параллельных задач, часто без зависимости между ними, например все, что требует распространения через пространственный параметр. Они затем могут быть распределены через доступные процессоры, принося существенное сокращение времени вычисления по сравнению с прямо направленным последовательным подходом. Если процессоры могут быть выбраны так, чтобы затребованные ими данные были локальными (локальность данных), то это в дальнейшем сокращает время вычисления (фактически, это общий принцип работы с крупными массивами данных – перенос вычисления на данные).

Существует несколько подходов к управлению этими вычислениями: *Condor* и *BOINC* (сеть распределенных вычислений) будут рассматривать общие работы в общих пулах машин, начиная от локальных ресурсов, посвященных этому процессу, до свободных циклов, удаляющих мусор из электронных ресурсов по всему миру (обычный сценарий для *BOINC*), – хотя данные постоянно переходят к вычислению с ними. Отметим, что графические процессоры предлагают все более по-

пулярную альтернативу кластерам центрального процессора с отдельными высококачественными чипами, обеспечивающими производительность высоких скоростей свыше ~1000 по сравнению с отдельными центральными процессорами, предполагающими соответствующую параллелизацию кодов. Фактически, кластеры графических процессоров производят массу грубых подсчетов, жизнеспособных в условиях современных алгоритмических подходов центрального процессора, например в функциях корреляции *n*-точек [15].

Модель распределенных вычислений *MapReduce* [16] и её эквивалент открытого источника (Hadoop) применяют иной подход — выражение работ в терминах двух стандартных операций — составление карты и уменьшение, примеры которых (картопостроители и преобразователи) используются для вычислительных ресурсов, содержащих данные для обработки (таким образом, гарантируя локальность данных). Картопостроитель преобразует свои входные данные (как ключевые, имеющие значение пары) в промежуточный набор различных (ключевых, имеющих значение) пар. Будучи собранными из всех картопостроителей, они записываются, и группа данных для каждого отдельного ключа посылается в преобразователь. Наконец, выходные данные преобразователей собираются и возвращаются как общий результат вычисления.

Не все вычисления выражаются в такой форме – те, которые требуют большого числа заявленной информации для обмена между картопостроителями, например, соотнесение общей обучающей последовательности с массой тонкоструктурной синхронизации может быть проблематичным, хотя те, которые включают итеративные процессы, могут быть выражены как цепочки задач модели *MapReduce*. Альтернативной моделью является применение текущего решения для вычисления, т.е. такого, которое требует только одного прохода через данные. Обычно они включают прирастающую (онлайн) формулировку вычислительного алгоритма, который совершенствуется с каждой новой позицией данных. Дальнейшая оптимизация возможна для определенных типов вычислений, таких как стохастический плавный спуск для некоторых типов машинного обучения. Очевидно, что для крупных массивов данных вычисления, основанные только на одном чтении данных, идеальны, и в некоторых случаях такие алгоритмы также применяются для параллелизации.

Подобно тому, как полифония лежит в сердце средневековой фуги с многочисленными голосами, сочетающимися для образования гармоничного целого, так параллелизация находится в центре фуги современных данных с тысячами центров и нитей, действующих в согласии с преобразованием огромных массивов данных в гармоничные представления нашего знания о космосе.

АСТРОЛОГИЯ ДАННЫХ

Принцип «как вверху, так и внизу» подкреплял средневековую веру в то, что модели на небесах отражают или даже предсказывают события, происходящие здесь на земле во всех сферах жизни, начиная от личного здоровья и до государственных дел, до триумфов и катастроф. Астрономия была наукой наблюдения этих моделей и их толкования, используя кодекс законов вавилонской и исламской мысли. Планы по созданию были широко выражены в астрономических расположениях звезд и планет, и мы могли предсказать их путем собственного изучения. Извлечение данных представляет собой «полуавтоматическое открытие моделей, связей, изменений, аномалий и статистически значимых структур и событий в данных» [17] и является основой астроинформатики.

Применение извлечения данных к массиву данных на самом деле имеет две первоначальные цели [18]: предсказание будущего поведения некоторых объектов на основе существующего поведения других объектов в данных (предсказание) и нахождение интерпретируемых человеком моделей, описывающих данные (описание) – интересно, что то же деление разграничивало судебную и естественную астрологию. Набор доступных методов для извлечения данных, возникший впервые в вычислительной технике (особенно – исследование искусственного интеллекта) и статистике, может быть затем рассмотрен как попадание в одну или более таких категорий, как классификация, регрессия, кластеризация, обобщение, зависимое моделирование, обнаружение изменения и отклонения (или выброса).

Однако процесс извлечения данных действительно распространяется за рамки просто случайного применения частного алгоритма. Представляющие интерес данные сначала должны быть собраны и тщательно подготовлены к анализу путём, например, нормализации, обработки пропущенных значений, сбора, выборки и т. д. Предположения и недостатки отдельного метода, который должен применяться, необходимо оценить, например следует определить конкретное число кластеров, и во многих случаях потребуется многочисленное применение алгоритма для полного их определения. Даже тогда результат должен подтверждаться либо путем повторного проведения анализа на подмножествах данных и/или путем использования некоторой особой меры качества. Наконец, процедура должна быть достаточно понятной, чтобы результаты могли быть интерпретированы и чтобы она (процедура) могла быть использована для дальнейших и более широких выборок данных.

Важным аспектом извлечения данных является включение соответствующего первоначального знания. Статистический логический вывод (см. раздел Логика данных) является одним из подходов к этому, но он строит свои аргументы на вероятностных моделях данных, а не на действительно наблюдаемых значениях. Таким образом, его интерпретации основываются на предположении, что модель служит хорошим описанием реальности, а не наблюдений. Свертывание знания в алгоритм извлечения данных, по крайней мере, означает, что любые интерпретации являются основанными на данных, даже если знание может быть произведено из модели. Исходя из семантических конструкций, таких как онтологии, могут быть определены метрики сходства, кодирующие степень, до которой два понятия разделяют информацию. Эти количественные меры концептуального сходства могут быть затем объединены в стандартные формулировки алгоритма извлечения данных с учетом извлечения знания из данных.

Из всех моделей, которые разглядели на небесах средневековые ученые, наиболее жизненной была *computus* (модель вычисления), которая позволяла определить дату Пасхи. Польза моделей, которую мы открыли в астрономических данных, приводит к открытию новых объектов, улучшенной обработке, выявлению объекта и классификации, а также к лучшим фотометрическим смещениям красных спектральных линий [19].

СХОЛАСТИЦИЗМ ДАННЫХ

Тривиум и *квадривиум* создали схоластическую культуру, в которой все явления, природные и искусственные, были предметом вопроса и символической интерпретации. Гуманитарные науки не только предоставляли необходимые навыки для раскрытия знания, спрятанного в создании, но и обеспечивали систему, в которой эти открытия могли добавляться и пониматься. В частности, свойства и отношения чисел, неизменных и бесконечных, были дорогой к пророческому откровению. Наше желание открыть скрытые работы универсума является неизменным, но мы больше не требуем, чтобы оно было таинственным. Научный метод, возникший из диалектической критики средневековья, основан на рациональной мысли и логике, имеющими дело с трудными данными и фактами, а не с ассоциациями и толковательной совместимостью.

Тем не менее, мы показали, как одни и те же темы проходят через наш современный подход. Имея наши огромные массивы данных, мы до сих пор обеспокоены структурами, которые применяем для представления нашего знания, хорошей и правильной их передачи, и тем, как мы можем их спроектировать и создать. Нам все еще необходимо понять, чем является то, что мы изучаем, и какие применять правила. И нам все еще надо знать, как искать значимые модели, которые мы хотим обнаружить. Только на этой основе мы можем надеяться на управление огромными объемами и сложностями данных, с которыми мы сталкиваемся.

Наконец, должна быть общая попытка, международная и междисциплинарная. Проблемы для астрономии являются теми же, что и для климатологии, геномики, для любой отрасли XXI в. Попытки таких организаций, как Международный союз виртуальной обсерватории [20], являются шагом в правильном направлении, но нам необходимо что-то истинно универсальное, обучающее на всех уровнях и во всех областях. Данные, подобные их средневековому аналогу (числу) должны быть объектом первого класса в нашем мировоззрении, и не только с технической точки зрения. С точки зрения будущего наше время должно рассматриваться как позиция, в которой мы взяли данные средневековья и положили начало правильной современной перспективе.

Благодарность. Автор выражает благодарность Норману Грею и Хелен Ангов за полезное сотрудничество и обсуждение этой статьи.

ЛИТЕРАТУРА

1. *Hey T., Tansley S., Tolle K.* The fourth paradigm. — Microsoft Research, Redmond, 2009.

2. *Kline M.* Mathematics in western culture. — Oxford University Press, Oxford, 1953.

3. *Gray N.* Data is a singular noun. — 2010. — http://purl.org/nxg/note/singular-data.

4. *Hogg D.W., Lang D.* Astronomical imaging: The theory of everything//American Institute of Physics Conference Series. — 2008. — Vol. 1082. — P. 331.

5. *Hogg D.W., Lang D.* Telescopes don't make catalogs! //EAS Publications Series. — 2011. — Vol. 45. — P. 351.

6. FITS image compression programs.— http://heasarc.gsfc.nasa.gov/ fitsio/fpack

7. *Drake A. at al.* First results from the Catalina real-time transient survey// ApJ. — 2009. — Vol. 696.

8. *Riess A. et al.* Observational evidence from supernovae for an accelereating universe and a cosmological constant//AJ.— 1998. — Vol. 116. — P. 1009.

9. *Kogge P. et al.* ExaScale Computing Study: Techonology challenges in achieving Exascale systems. — 2008. — doi: 10.1.1.165.6676.

10. *Fluke C.J., Barnes D.G., Barsdell B.R., Hassan A.H.* Astrophysical supercomputing with GPUs: Critical decisions for early adopters//PASA.— 2011. — Vol. 28. — P. 15-27.

11. *Ghemaway S., Gobioff H., Leung S.T.* The Google File System. — 2003. — http://labs.google.com/papers/gfs.html.

12. *Wiley K., Connolly A., Gardner J., Krughoff S., Balazinska M., Howe B., Kwon Y., Bu Y.* Astronomy in the cloud: Using MapReduce for image co-additon//PASP.— 2011. — Vol. 123. — P. 366-380.

13. *Chang F. et al.* BigTable: A distributed storage system for structured data. — 2006. — http://labs.google.com/ papers/bigtable.html.

14. *Gray J., Hey T.* In search of petabyte databases. — 2001. — http://www.research.microsoft.com/~Gray/talks/ In%20Search%20of%20Petabyte%20Databases.

15. *Tian H.J., Neyrinck M.C., Budavari T., Szalay A.S.* Redshift-space enhancement of line-of-sight baryon acoustic oscillations in the sloan digital sky survey main-galaxy sample//ApJ.— 2011.— Vol. 728.— P. 34.

16. *Dean J., Ghemawat S.* MApReduce: Simplified data processing on large clusters. — 2004. — http://labs.google.com/papers/mapreduce.html.

17. IVOA Knowledge Discovery in Databases. — http://www.ivoa.net/cgi-bin/twiki/bin/view/IVOA/IvoaKDD.

18. *Fayyad U., Piatetsky-Shapiro G., Smyth P.* From data mining to knowledge discovery in databases//AI Magazine. — 1996. — Vol. 17. — P. 37-54.

19. Data mining examples. — http://www.ivoa.net/cgi-bin/twiki/bin/view/IVOA/IvoaKDDguideScience.

20. International Virtual Observatory Alliance (IVOA). — http://www.ivoa.net

Влияние цитирования статей, опубликованных в шести продуктивных странах: сравнение по странам, основанное на данных InCites*

Луц БОРНМАН
(Lutz BORNMANN)

Общество Макса Планка,
г. Мюнхен, Германия

Лут ЛЕЙДЕСДОРФ
(Loet LEYDESDORFF)

Амстердамская школа исследования коммуникации, Амстердамский университет, г. Амстердам, Нидерланды

Используя инструмент InCites издательства Thomson Reuters, данное исследование сравнивает нормализованные значения влияния цитирования, вычисленные для Китая, Японии, Франции, Германии, США и Великобритании за 1981 – 2010 гг. Значения влияния цитирования нормализованы для четырех предметных областей: естественные науки, инженерия и технология, медицинские науки и здравоохранение, сельскохозяйственные науки. Результаты показывают тенденцию роста в значениях влияния цитирования для Франции, Великобритании и особенно для Германии на протяжении последних тридцати лет во всех предметных областях. Влияние цитирования статей из Китая находится все еще на относительно низком уровне (по большей части ниже среднего мирового значения), но страна следует линии тенденции роста. США демонстрируют относительно стабильную модель высоких значений влияния цитирования на протяжении анализируемых лет. С небольшими различиями влияния между годами публикации тенденция США показывает рост в области инженерии и технологии, однако происходит снижение в медицинских науках и здравоохранении, а также в сельскохозяйственных науках. Подобно США, Япония следует тенденциям роста и снижения в различных предметных областях, но изменения на протяжении рассматриваемых лет являются небольшими. В большинстве ряда лет статьи из Японии дают результат ниже или приблизительно на уровне среднего мирового значения в каждой предметной области.

ВВЕДЕНИЕ

Исследования, концентрирующиеся на международных сравнительных характеристиках стран или территорий, имеют давнюю традицию. Лейдесдорф и Вагнер утверждают, что «США пока превосходят все другие страны по отношениям высокоцитируемых статей и ссылки/публикация и являются более успешными, чем Европейский союз (ЕС), в координировании своих научных усилий в стратегически приоритетных областях, таких как нанотехнология» [1, с. 23]. Отчет Национального научного управления США, содержащий показатели по науке и инженерии, сообщает о достижениях в международной и американской науке и технологии, основанных на полной базе данных. Этот отчет показывает, что общая доля опубликованных статей ученых в ЕС и США «постепенно снизилась с 69 % в 1995 г. до 58 % - в 2009 г. Немногим более чем за десятилетие мировая доля статей из Азии возросла с 14 % до 24 %, будучи возглавляемой 16 % среднего ежегодного роста Китая, доминирующего в мировом производстве статей» [2, с. 0 – 10]. Что касается влияния цитирования, то анализы Национального научного управления США [2] обнаруживают, что «статьи США продолжают иметь самые высокие темпы цитирования во всех широких областях науки и инженерии» [2, с. 5 – 43]. На превосходном уровне США опубликовали на 76 % больше статей, чем ожидалось среди 1 % цитируемых статей верхнего ранга в 2010 г.; ученые из ЕС опубликовали 1 % статей верхнего ранга, на 7 % меньше, чем предполагалось.

Два недавно опубликованных отчета, хотя и основанные на разных базах данных (Web of Science – WoS издательства Thomson Reuters и Scopus издательства Elsevier), пришли к одному и тому же выводу: по сравнению с дру-

* Перевод Bornmann L., Leydesdorff L. Citation impact of papers published from six prolific countries: A national comparison based on incites data. – http://arxiv.org/abs/1205.0680

гими высокоразвитыми странами Великобритания занимает ведущее место (первое и второе, соответственно) по ссылкам на статью [3]. Согласно работе [4], «данные Великобритании находятся на поднимающейся траектории, тогда как научная база США имеет ровное направление в своей характеристике и, по некоторым вычислениям, в настоящее время находится в стадии снижения, … сейчас Великобритания догоняет США по средним значениям и … заметно опережает Германию и Францию. Япония достаточно проигрывает своим конкурентам по странам Большой семерки (G 7), и ее среднее влияние сейчас схоже с характеристикой Китая» [4, с. 8]. В то время как Адамс [4] использовал данные WoS, издательство Elsevier [5] приводит следующие результаты, основанные на данных Scopus: влияние цитирования взвешенной области в Великобритании… значительно выше мировой отметки, но немного ниже, чем таковое США, обеспечивающее второе место в G 8 и в конкурирующей группе, но с темпом роста в 1,1 % по сравнению с США – 0,5 % в год с 2006 г.» [5, с. 28].

В данном исследовании анализы влияния цитирования Адамса [4] и Маршалла и Тревиса [3] расширены путем: 1) рассмотрения более длинного публикационного интервала (1981 – 2010 гг. вместо 1991 – 2010 гг.), 2) дифференциации в четырех широких предметных областях (вместо анализа по всем областям и некоторым выборочным дисциплинам) и 3) использования статистических процедур для исследования тенденций влияния цитирования на протяжении рассматриваемого ряда лет. Согласно [3, 4], анализы результатов данного исследования относятся к шести наиболее развитым странам: Китай, Япония, Франция, Германия, США и Великобритания. Адамс [4] аргументирует выбор этих шести стран следующим образом: они «в настоящее время отвечают за две трети общего числа научных публикаций, проиндексированных в Web of Knowledge^{SM} издательства Thomson Reuters. Франция и Германия являются основными партнерами Великобритании в европейской научной сфере. США и Япония являются ее основными коллегами по сотрудничеству в других регионах. Китай – это новый динамичный фактор в глобальной географии науки, как мы сообщали ранее. Его быстрый подъем демонстрирует другую модель в научной политике и требует повторного анализа несомненных фактов, которые поддерживали трансатлантическую ось в течение последней половины века» [4, с. 3].

Данное исследование основывается на данных из сравнительно нового средства InCites издательства Thomson Reuters, которое облегчает национальные сравнения по длинным временным периодам с использованием нормализованных значений влияния цитирования.

МЕТОДЫ

Средство InCites

InCites (Thomson Reuters; http://incites.thomsonreuters.com/) является сетевым средством научной оценки, позволяющим оценивать продуктивность и влияние цитирования учреждений (институтов, организаций) и стран. Модуль глобальных сравнений обеспечивает показатели цитирования из WoS для оценки научной продукции учреждений и стран. Показатели получены из набора данных о 22 млн статей WoS за 1981 – 2010 гг. Показатели специальных сравнений стран созданы на основе критериев адресов с использованием метода полного подсчета (например, адреса авторов, опубликовавших статьи). Издательство Thomson Reuters использует полный подсчет: подсчеты не оцениваются (взвешиваются) по числу авторов или числу адресов.

Показатели, характеризующие страны, могут быть получены как отчет сравнения по странам в формате Excel. В качестве схемы предметных областей для данного исследования использовались основные категории Организации экономического сотрудничества и развития (OECD) [6]. (InCites обеспечивает шесть дальнейших схем, например 22 предметные области, предоставляемые издательством Thomson Reuters в Essential Science Indicators. Таблица соответствия между категориями OECD и предметными категориями WoS также представлена.) По сравнению с другими схемами схема OECD дает возможность использовать шесть широких предметных категорий данных WoS: 1) естественные науки, 2) инженерия и технология, 3) медицинские науки и здравоохранение, 4) сельскохозяйственные науки, 5) общественные науки и 6) гуманитарные науки. Каждая предметная категория включает подчиненные области. Например, сельскохозяйственные науки включают следующие входящие в нее предметные области: 1) сельское хозяйство, лесное хозяйство и рыболовство; 2) животноводство и молочное хозяйство; 3) ветеринария; 4) сельскохозяйственная биотехнология и 5) другие сельскохозяйственные науки.

Для каждой широкой предметной категории данные по странам [7] были получены в виде листа Excel и импортированы в Stata [8] для статистического анализа. Однако данные для общественных наук и гуманитарных наук не были включены в данный анализ. Согласно [9], «вряд ли можно ожидать, что ученые в области гуманитарных и общественных наук примут показатели производительности, используемые в естественных науках в качестве имеющих силу в их собственной области. В результате необходимы альтернативные методы. Успех любых оценочных процедур и инструментов зависит от их принятия соответствующими учеными [9, с. 5].

Показатели цитирования и статистические процедуры

Чтобы иметь надежные значения влияния цитирования в данном исследовании, в статистический анализ предметной области включены только те годы публикации, в которые имелось, по меньшей мере, 100 статей. Например, данные Китая по сельскохозяйственным наукам не были включены за 1981 – 1989 гг. Следуя представлению [3, 4], мы показываем далее нормализованные значения влияния цитирования для шести различных стран. Издательство Thomson Reuters вычисляет среднюю частоту цитирования для множества публикаций страны и затем делит его на среднюю частоту всех публикаций в этой предметной области. Значение 1 для определенной страны в определенной предметной области означает, что влияние цитирования статей, опубликованных учеными в этой стране, не больше и не меньше, чем среднее влияние статей в этой предметной области*. Если добавить это значение, то получится, например, что 1,2 соответствующих статей цитировались в среднем на 20 % выше среднего в этой области.

Для каждой страны в анализе определенной предметной области вычисляется упорядоченный коэффициент Спирмана (ρ) для корреляции между годом публикации и влиянием цитирования [12]. Эти коэффициенты поддерживают интерпретацию результатов тенденции вне визуального взгляда на эти кривые. Стати-

* Поскольку нормализация проводится для одной категории в каждом случае, деноминатор является постоянным и, следовательно, вопрос различия между «нормами средних значений» и «средними значениями норм» не играет роли в этой нормализации [10, 11].

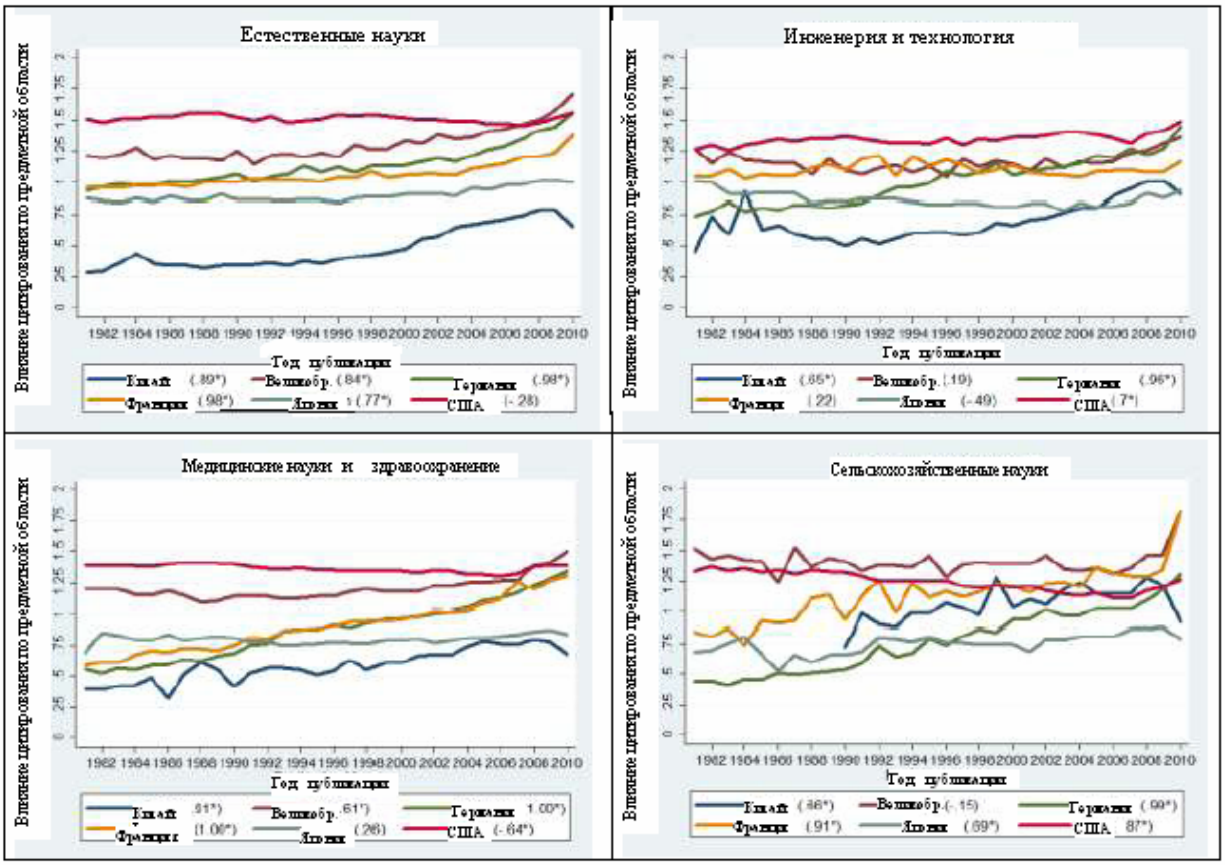
стически значимый коэффициент (ρ <0,01, протестированный с двух сторон; критическое абсолютное значение коэффициента - 0, 478) [13] показывает, может ли «истинный» коэффициент (коэффициент совокупности), на самом деле быть выше 0. Поскольку коэффициенты вычисляются на основе ранговых данных, они не учитывают размера различий влияния между двумя годами публикации, а учитывают только существование различия. Это означает, что страна с тенденцией роста значений влияния может иметь еще меньший абсолютный рост между 1981 – 2010 гг., чем другая страна с (более или менее) произвольным распределением значений, но с некоторыми определенными повышениями влияния за временной период.

Для измерения изменчивости значений влияния цитирования за годы публикации вычисляется стандартное отклонение для каждой страны и каждой категории области. Стандартное отклонение показывает степень отклонений от среднего влияния цитирования страны за все публикационные годы. Относительно небольшое стандартное отклонение, например, показывает, что значения влияния не отклоняются в большой степени от среднего значения на протяжении лет.

Результаты

На рисунке представлены значения влияния цитирования для шести стран, вычисленные соответственно для четырех областей. Необходимо обратить внимание на два момента в интерпретации результатов: 1. Большие различия влияния цитирования между двумя последующими годами для одной страны скорее могут быть по причине более низкого числа статей, чем значимых различий характеристик. Поскольку авторы из США публикуют больше статей, чем все другие страны, то изменение является соответственно низким. 2. Чем больше интервал цитирования, тем более надежным является подсчет характеристики статьи [14]. Следовательно, самые последние публикационные годы на рисунке должны интерпретироваться с должным вниманием (см., например, подъемы в 2010 г. в значениях влияния цитирования для сельскохозяйственных наук).

Если мы сравним результаты различных предметных областей на этом рисунке с рисунками, опубликованными Маршаллом и Тревисом [3] и Адамсом [4] по всем предметным областям, то самым большим будет соответствие для естественных наук. В этой предметной области мы видим подобную картину с резким ростом во влиянии статей из Великобритании за последние четыре года. Великобритания является единственной страной, которая превосходит по показателям США в течение последних тридцати лет. Интересно наблюдать за естественными науками, по которым все страны, за исключением США, имеют очень высокие коэффициенты корреляции (Китай: ρ = 0,89, Великобритания: ρ = 0,84, Германия: ρ = 0,98, Франция: ρ = 0,98, Япония: ρ = 0,77), указывающие на тенденцию возрастания значений влияния цитирования.



Значения влияния цитирования по шести странам (Китай, Великобритания, Германия, Франция, Япония и США), вычисленные относительно четырех предметных областей (естественные науки; инженерия и технология; медицинские науки и здравоохранение и сельскохозяйственные науки). Для каждой страны дается упорядоченный коэффициент Спирмана для корреляции между годом публикации и влиянием цитирования; звездочка означает статистически значимый коэффициент («истинный» коэффициент, коэффициент совокупности, вероятно, должен быть выше 0). Высокий коэффициент корреляции показывает тенденцию роста или снижения в значениях влияния цитирования.

Источник: InCites™ Thomson Reuters [7].

Эта тенденция особенно заметна в европейских странах (Франция, Германия и Великобритания). На низком уровне представления данная тенденция также видна для Китая, но она снижалась в течение двух последних лет. Исследования в последующие несколько лет покажут, будет эта тенденция продолжаться или нет.

Для инженерии и технологии на рисунке наблюдается иная модель, нежели для естественных наук. Не только Германия (ρ = 0,96) и Китай (ρ = 0,65), но также США (ρ = 0,7) показывают тенденцию роста влияния цитирования на относительно высоком уровне (эти три страны имеют самые высокие коэффициенты корреляции на диаграмме). Противоположная тенденция наблюдается для Японии (ρ = 0,49, но этот коэффициент не является статистически значимым). В следующей диаграмме рисунка, медицинские науки и науки здравоохранения, четко просматривается тенденция роста для Германии (ρ = 1) и Франции (ρ = 1), но тенденция снижения для США (ρ = - 0,64). Однако, если тенденция снижения США происходит на уровне высокого влияния в течение последних тридцати лет, то Германия и Франция улучшили свое влияние с очень низкого уровня представления в 1980-х гг. до очень высокого уровня (на 25 % выше среднего по предметной области) за последние два года. Сельскохозяйственные науки – как последняя предметная область, рассматриваемая на рисунке, - характеризуется тенденциями роста Китая (ρ = 0,66), Франции (ρ = 0,91) и Германии (ρ = 0,99), но тенденцией снижения США (ρ = - 0,87). За весь временной период Великобритания действует на очень высоком уровне, по крайней мере, на 25 % выше среднего по предметной области.

В таблице приводятся минимальные, максимальные, средние и стандартные отклонения значений влияния цитирования для различных предметных областей и стран. Хотя США имеют в большинстве предметных областей на рисунке статистически значимую тенденцию (возрастание или снижение), стандартные отклонения (SD) во всех предметных областях очень небольшие (SD < 0, 1). Это означает, что существует только низкий уровень изменений между годами публикации. Другая ситуация отмечается для Германии (SD между 0,15 и 0,25) и Китая (SD между 0,13 и 0,16). Эти страны имеют высокий уровень изменения и тенденцию динамического роста (как показано на рисунке).

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

В большом соответствии с другими исследованиями (описанными во Введении) данное исследование показывает тенденцию роста нормализованных значений влияния цитирования для Франции, Великобритании и особенно для Германии за последние тридцать лет в четырех предметных областях. Для Франции и Германии это также может быть эффектом интернационализации. Ученые обеих стран в значительной степени публиковались на английском языке. Таким образом, измеряется не только научная характеристика, но также

Минимальное, максимальное, среднее и стандартное отклонение значений влияния цитирования по предметным областям и странам (*N* - число публикационных лет, включенных в данный анализ)

	Китай	Великобритания	Германия	Франция	Япония	США
Естественные науки						
N	30	30	30	30	30	30
Минимальное	0,29	1,14	0,93	0,95	0,85	1,45
Максимальное	0,78	1,71	1,55	1,38	1,01	1,55
Среднее	0,47	1,30	1,13	1,06	0,90	1,51
Стандартное отклонение	0,16	0,13	0,15	0,10	0,05	0,03
Инженерия и технология						
N	30	30	30	30	30	30
Минимальное	0,45	1,05	0,73	1,03	0,77	1,25
Максимальное	1,01	1,35	1,45	1,21	1,01	1,48
Среднее	0,69	1,16	1,00	1,10	0,87	1,34
Стандартное отклонение	0,15	0,07	0,19	0,05	0,05	0,05
Медицинские науки и здравоохранение						
N	30	30	30	30	30	30
Минимальное	0,33	1,09	0,53	0,59	0,69	1,30
Максимальное	0,79	1,50	1,34	1,30	0,87	1,41
Среднее	0,58	1,20	0,87	0,90	0,80	1,36
Стандартное отклонение	0,13	0,09	0,24	0,20	0,04	0,03
Сельскохозяйственные науки						
N	21	30	30	30	30	30
Минимальное	0,72	1,24	0,42	0,73	0,53	1,11
Максимальное	1,28	1,83	1,30	1,83	0,88	1,36
Среднее	1,06	1,41	0,76	1,13	0,7	1,24
Стандартное отклонение	0,14	0,10	0,25	0,21	0,08	0,08

сдвиг производительности. Влияние цитирования статей из Китая все еще находится на относительно низком уровне (в большинстве случаев ниже мирового среднего), но страна следует линии тенденции роста. США показывают сравнительно устойчивую модель высоких значений влияния цитирования за указанные годы. С небольшими различиями влияния между годами публикации тенденция США стремится к росту в сфере инженерии и технологии, но к снижению в области медицины и здравоохранения, а также в сельскохозяйственных науках. По аналогии с США Япония следует тенденциям роста и снижения в различных предметных областях, но изменения на протяжении рассматриваемых лет являются очень небольшими. На протяжении большинства лет статьи из Японии давали картину ниже или приблизительно на уровне мирового среднего значения в предметной области.

В качестве первого недостатка данного исследования необходимо упомянуть использование средних арифметических темпов (частот) цитирований. Существует опасность концентрации на критериях центральной тенденции: несмотря на необычно распределяемые данные цитирования, арифметическое значение может дать искаженную картину вида распределения [15], «а это достаточно грубая статистика» [16, с.1]. Поскольку распределение частот цитирования обычно является прямо противоположным распределению в соответствии со степенной зависимостью, среднеарифметические темпы цитирований в основном показывают, где статьи с высокими темпами цитирования должны быть обнаружены [17].

Особенно в библиометрическом анализе использование процентных ранговых значений в целях оценки является очень полезным, поскольку не надо делать никаких предположений о распределении ссылок [18]. Хотя некоторые сравнительные исследования по странам дают дополнительно результаты, основанные на процентных отношениях [2], обычно наиболее используемым подходом является представление (нормализованных) средних темпов (частот) цитирования. Поскольку InCites не обеспечивает процентные отношения в модуле глобального сравнения, данное исследование следовало этому подходу. Для будущих сравнений по странам очень желательны исследования, основанные на процентных отношениях (см., например, новое Лейденское ранжирование – http://www.leidenranking.com/ для сравнения учреждений) [19].

Второй недостаток нашего исследования касается использования четырех широких областей для нормализации результатов. Осуществление нормализации на уровне этих четырех широких областей означает, что высокое влияние цитирования страны в определенной области может быть следствием того факта, что большая часть деятельности страны в этой области имеет место в подобластях со сравнительно высокой плотностью цитирования. Таким образом, различия во влиянии цитирования могут просто отражать различия в научной деятельности стран (а не различия в актуальном научном влиянии страны). Также рост (или снижение) влияния цитирования страны в определенной области может быть последствием перехода научной деятельности страны от областей с более низкой (высокой) плотностью цитирования к подобластям с более высокой (низкой) плотностью цитирования.

ЗАКЛЮЧЕНИЕ

Что касается модуля глобальных сравнений, InCites предлагает уникальную возможность изучать нормализованное влияние цитирования стран. Хотя подход имеет свои недостатки: 1) рассмотрение длинных публикационных периодов (1981 – 2010 гг.), 2) дифференциация в (широких) предметных областях и 3) использование статистических процедур, он позволяет осуществлять проницательное исследование национальных тенденций влияния цитирования на протяжении определенных лет.

ЛИТЕРАТУРА

1. *Leydesdorff L., Wagner C.* Is the United States losing ground in science? A global perspective on the world science system // Scientometrics. – 2009. – Vol. 78, No. 1. – P. 23-26/ - doi: DOI 10.1007/s11192-008-1830-4.

2. *National Science Board* // Science and engineering indicators 2012. – Arlington, VA, USA: National Science Foundation (NSB 12-01), 2012.

3. *Marschall E., Travis J.* UK scientific papers rank first in citations // Science. - 2011. – Vol. 334, No. 6055. – P. 443.

4. *Adams J.* Global research report: United Kingdom. – Leeds, UK: Evidence, 2010.

5. *Elsevier.* International comparative performance of the UK research base – 2011. – London, UK: Department of Business, Innovation and Skills (BIS), 2011.

6. *Organisation for Economic Co-operation and Development.* Revised field of science and technology (FOS) classification in the Frascati manual. – Paris, France: Working Party of National Experts on Science and Technology Indicators, Organisation for Economic Co-operation and Development (OECD), 2007.

7. *InCites™ Thomson Reuters* // Report created; 02. 2012. Data Processed Dec 31, 2010. Data source: Web of Science. This data is reproduces under a license from Thomson Reuters, 2011.

8. *StataCorp* // Stata statistical software: release 12. – College Station, TX, USA: Stata Corporation, 2011.

9. *Blockmans W., Thomassen J.* Judging research on its merits. – Amsterdam, The Netherlands: Royal Netherlands Academy of Arts and Sciences, 2005.

10. *Egghe L.* Averages of ratios compared to ratios of averages: Mathematical results // Journal of Informetrics. – 2012. – Vol. 6, No. 2. – P. 307-317. – doi: 10.1016/j.joi.2011.12.007.

11. *Gingras Y., Larivière V.* There are neither "king" nor "crown" in scientometrics: Comments on a supposed "alternative" method of normalization // Journal of Informetrics. – 2011. – Vol. 5, No. 1. – P. 226-227. – doi: 10.1016/j.joi.2010.10.005.

12. *Sheskin D.* Handbook of parametric and nonparametric statistical procedures (4th ed.). – Boca Raton, FL, USA: Chapman & Hall/CRC, 2007.

13. *Glass G.V., Stanley J.C.* Statistical methods in education and psychology. – Englewood Cliffs, NJ, USA: Prentice Hall, 1970.

14. *Research Evaluation and Policy Project* // Quantitative indicators for research assessment – a literature review (REPP discussion paper 05/1). – Canberra, Australia: research Evaluation and Policy Project, Research School of Social Sciences, The Australian National University, 2005.

15. *Bornmann L., Mutz R., Neuhaus C., Daniel H.-D.* Use of citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results // Ethics in Science and Environmental Politics. – 2008. - Vol. 8. – P. 93-102. – doi: 10.3354/esep00084.

16. *Adler R., Ewing J., Taylor P., Hall P.G.* A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS) // Statistical Science. – 2009. – Vol. 24, No. 1. – P. 1-28. – doi: 10.1214/09-sts285.

17. *Bornmann L., Mutz R., Marx W., Schier H., Daniel H.-D.* A multilevel modelling approach to investigation the predictive validity of editorial decisions: Do the editors of a high-profile journal select manuscripts that are highly cited after publication? // Journal of the Royal Statistical Society – Series A (Statistics in Society). – 2011. – Vol. 174, No. 4. – doi:10.1111/j.1467-985X.2011.00689.x.

18. *Leydesdorff L., Bornmann L., Mutz R., Opthof T.* Turning the tables in citation analysis one more time: Principles for comparing sets of documents //Journal of the American Society for Information Science and Technology. – 2011. – Vol. 62, No. 7. – P. 1370-1381.

19. *Bornmann L., de Moya Anegón F., Leydesdorff L.* The new excellence indicator in the world report of the SCImago institutions rankings 2011 // Journal of Informetrics. – 2012. – Vol. 6, No. 2. – P. 333-335.

Ослабление взаимосвязи между фактором влияния и цитированиями статей в цифровой век*

Жорж А. ЛОЗАНО
(George A. LOZANO)

Обсерватория наук и технологий, Межуниверситетский центр исследований по науке и технологии, Университет Квебека в Монреале, г. Монреаль, Канада

Винсент ЛАРИВЬЕР
(Vincent LARIVIERE)

Школа библиотековедения и информатики, Монреальский университет, г. Монреаль, Канада

Ив ЖИНГРА
(Yves GINGRAS)

Обсерватория наук и технологий, Межуниверситетский центр исследований по науке и технологии, Университет Квебека в Монреале, г. Монреаль, Канада

Исторически статьи физически связаны с журналом, в котором они были опубликованы, но в электронный век статьи доступныпоодиночке и больше не привязаны к соответствующему журналу. Следовательно, теперь статьи можно читать и цитировать на основе их собственных достоинств независимо от физической доступности, репутации или фактора влияния журнала. Мы сравниваем силу взаимосвязи между факторами влияния журналов и действительными ссылками, полученными их соответствующими статьями в 1902 – 2009 гг. В течение XX в. частота цитирования статей была в большей степени связана с факторами влияния их соответствующих журналов. Однако с 1990 г., с наступлением цифрового века, сила взаимосвязи между факторами влияния и цитированиями статей снижается. Раньше всех снижение началось в физике, области, которая смогла быстрее сделать переход в электронную сферу. Кроме того, с 1990 г. доля высокоцитируемых статей, происходящих из высокоцитируемых журналов, снизилась и, соответственно, доля высокоцитируемых статей, не происходящих из высокоцитируемых журналов, увеличилась. В случае, если эта тенденция продолжится, то она может положить конец использованию фактора влияния как способа оценки качества журналов, статей и ученых.

ВВЕДЕНИЕ

Изначально фактор влияния появился в 1960-х гг., чтобы ориентировать академические библиотеки в их приобретениях журналов [1]. Хотя в последнее время появился ряд других основанных на цитировании видов критериев влияния журнала, таких как SCImago Journal Rank [2], собственный фактор (Eigenfactor) [3] и нормализованное влияние источника на статью [4], двухлетний фактор влияния, составленный издательством Thomson Reuters, все еще наиболее широко используется. Фактор влияния определенного журнала за определенный год определен как средняя частота цитирования в течение этого определенного года статей, опубликованных в этом журнале за предыдущие два года. Например, фактор влияния журнала за 2011 г. учитывает ссылки, полученные в 2011 г. статьями, опубликованными в этом журнале за 2009 – 2010 гг. Таким образом, интервал ссылок для отдельных статей колеблется от одного года до почти трех лет, в среднем – два года.

За несколько прошедших десятилетий факторы влияния постепенно укоренились в коллективном сознании ученых и стали самоусиливающимися критериями качества журналов, статей и их авторов. Ученые теперь учитывают факторы влияния при выборе журналов для своих публикаций; редакторы журналов формулируют стратегии, разработанные исключительно для улучшения своих факторов влияния, а издатели рекламируют свои факторы влияния на своих сетевых сайтах. Факторы влияния часто используются как заменитель действительного числа ссылок, которое в конечном счете может получить недавно опубликованная статья. Такое полномочие может быть частично оправдано, учитывая, что, независимо от качества статьи, фактор влияния журнала позитивно связан со ссылками, полученными его статьями [5]. С начала 1990-х гг., когда данные цитирования стали доступными в электронном виде, интерес к фактору влияния и его использование повысились, а количество научных статей о факторе влияния экспоненциально выросло [1].

* Перевод Lozano G.A., Larivière V., Gingras Y. The weakening relationship between the impact factor and papers' citations in the digital age. – http://arxiv.org/abs/1205.4328

Цифровой век также породил еще одно изменение. Со времени создания в 1665 г. *Journal des Sçavans* и *Philosophical Transactions of the Royal Society,* которые считаются двумя первыми научными периодическими изданиями, ученые в основном читали фактически существующие печатные журналы, поэтому статьи, опубликованные в узкопрофильных журналах с высоким обращением, имели больший шанс быть прочитанными и процитированными, чем статьи, опубликованные в менее широко доступных журналах. Теперь, когда научная информация распространяется электронным способом, менее вероятно, что ученые читают журналы целиком, вместо этого они проводят поиски электронной литературы по определенным темам и находят конкретные статьи из широкого разнообразия журналов. Следовательно, пока журнал находится в списке основных баз данных (например, Web of Science, Scopus или Google Scholar), а статьи доступны в электронном виде, они могут читаться и цитироваться на основе их собственных достоинств, независимо от физической доступности, репутации или фактора влияния журнала.

Таким образом, до наступления электронного века, частота цитирования любой статьи и фактор влияния ее журнала взаимно усиливали друг друга. Фактор влияния журнала основывался (и все еще продолжает основываться) на частоте цитирования его отдельных статей, а на частоту цитирования любой отдельной статьи влияли распространение журнала и его доступность, которые зависели от его фактора влияния. Теперь первое все еще остается справедливым, но если новые практики поиска литературы и использования ограничивают эффект фактора влияния журнала на частоту цитирования статьи, то корреляция между частотой цитирования статьи и фактором влияния должна со временем снижаться. Кроме того, доля высокоцитируемых статей из журналов с самым высоким фактором влияния должна со временем сокращаться. Здесь мы изучаем, действительно ли это так, и рассматриваем последствия продолжающегося использования фактора влияния для будущего научного издательства. Приводятся данные за 1900 – 2011 гг. по трем группам дисциплин: естественные и медицинские науки, взятые вместе, физика и общественные науки.

МЕТОДЫ

Мы использовали данные базы данныхWeb of Science (WoS) издательства Thomson Reuters за 1900 – 2011 гг., охватывающие все области естественных наук, медицинских наук и общественных наук. Набор данных охватывает Century of Science и Century of Social Sciences datasets за 1900 – 1944 гг. и Science Citation Index (расширенная версия), Social Sciences Index и Arts and Humanities Citation Index за период 1945 – 2011 гг. Классификация журналов по дисциплинам, используемая для естественных и медицинских наук вообще, и физики и общественных наук, в частности, является адаптацией классификации, используемой Национальным научным фондом США, которая категоризирует каждый журнал только в одной дисциплине и специальности.

Набор данных включал 25 569 603 статьи по естественным и медицинским наукам, 3 211 026 статей по физике и 879 494 статьи по общественным наукам. Общее число проанализированных цитируемых ссылок составляло 819 369 970. Статьи по гуманитарным наукам были исключены из анализа из-за их длинных интервалов цитирования и высоких показателей нецитируемости [5], однако ссылки на гуманитарные журналы были включены. Чтобы быть включенными в анализ, статьи должны были быть опубликованы в журнале, для которого должен быть вычислен фактор влияния. Однако ссылки, сделанные исключенными статьями, рассматривались как ссылки на другие статьи, включенные в анализ. Ряд журналов за некоторые годы не имели фактора влияния либо потому, что их статьи не получили никаких ссылок за двухлетний интервал цитирования, либо потому, что два года могли пройти прежде, чем новые журналы получили их первые ежегодные факторы влияния.

Учитывая, что издательство Thomson Reuters не составляет факторы влияния за весь изучаемый период и что точный метод, с помощью которого оно вычисляет факторы влияния, не является полностью ясным и, следовательно, воспроизводимым [6,7], фактор влияния каждого журнала, входящего в базу данных, был вычислен заново. В факторе влияния издательства Thomson Reuters некоторые типы публикаций используются для подсчета ссылок (номинатор), но сами не считаются «статьями» (деноминатор). Здесь фактор влияния был вычислен тем же способом, что и фактор влияния издательства Thomson Reuters, за исключением того, что: 1) эта асимметрия между номинатором и деноминатором была исправлена; за исключением, как отмечалось выше, в отеношении статей по гуманитарным наукам, если статьи были подсчитаны, их ссылки также были подсчитаны, 2) ссылки на отдельные статьи были подсчитаны за два полных года, следующие за их соответствующим годом публикации. Следовательно, данные о факторе влияния не были доступны за первые два года, и полные двухлетние интервалы цитирования были не возможны за последние два года, оставляя полный набор данных как факторов влияния, так и частоты цитирования за 1902 – 2009 гг. Большая доля статей осталась полностью нецитируемой в начале периода [5]; для уменьшения их веса во взаимосвязи фактор влияния / цитирование (ссылки) также были проведены анализы, исключающие нецитируемые статьи при вычислении как частоты цитирования, так и факторов влияния.

Использовались два показателя для измерения силы взаимосвязи между фактором влияния журналов и цитированиями статей. Первым показателем был коэффициент смешанной корреляции (r^2). Каждой статье приписывался фактор влияния журнала, в котором она была опубликована, и ссылки, которые она получила за два года, последовавшие за годом ее публикации, а r^2 между двумя рядами переменных подсчитывался за каждый год. Второй показатель – это ежегодная доля самых цитируемых статей, опубликованных в наиболее высокоцитируемых журналах, и ежегодная доля самых высокоцитируемых статей, не опубликованных в наиболее высокоцитируемых журналах.

РЕЗУЛЬТАТЫ

На рис. 1 - 3 представлены результаты r^2 между фактором влияния и цитированиями статей за 1902 – 2009 гг. для всех дисциплин естественных на-

ук и медицинских наук, взятых вместе (рис. 1), физики (рис. 2) и общественных наук (рис. 3). В описательных целях в случаях, где имел место явный разрыв, эти данные делились на периоды 1902 – 1958, 1959 – 1990 и 1991 – 2009 гг. Для медицинских и естественных наук (рис. 1) существовало увеличение корреляции между фактором влияния и частотой цитирования статей с 1902 г. до конца 1990-х гг. Сила взаимосвязи между фактором влияния и цитированиями стабильно не увеличивалась в течение XX в. Два отклонения имели место после двух мировых войн, вероятно, как результат изменений в научной системе. Наиболее интересным и иным в противовес общей тенденции, существовавшей на протяжении большей части XX в., поскольку научная информация начала распространяться в электронном виде примерно в 1990 г., является то, что взаимосвязь между фактором влияния и частотой цитирования начала ослабевать.

Такие же анализы были выполнены на двух дисциплинах, которые понимались как противоположные концы спектра в зависимости от того, насколько быстро они сделали переход в электронную область: физика (рис. 2) и общественные науки (рис. 3). Учитывая небольшой размер выборки, отклонение значений r^2 между фактором влияния и частотой цитирования статей больше, но в обоих случаях произошло снижение за последние два десятилетия. Хотя снижение не является в значительной степени различным в двух дисциплинах, в физике оно, кажется, началось раньше, к концу 1980-х гг. (рис.2).

Хотя это не показано, но все результаты являются схожими и выводы такими же, как при использовании r Пирсона и ранговой корреляции Спирмана в качестве показателей. Все анализы также были выполнены с исключением нецитируемых статей, как на уровне статей, так и на уровне вычисления факторов влияния журналов. Когда нецитируемые статьи исключаются, то возникает более ясная тенденция с небольшими изменениями, но сила взаимосвязи между фактором влияния и цитированиями остается в пределах той же величины. Таким образом, исключение нецитируемых статей не обеспечивает более сильную взаимосвязь между фактором влияния и цитированиями.

На рис. 4 и 5 представлен дополнительный показатель взаимосвязи между фактором влияния и цитированиями статей для всех дисциплин в естественных и медицинских науках: доля статей, которые составляют верхние (высшего ранга) 10 % (и 5 %) как наиболее цитируемых, так и публикуемых в ведущих 10 % (и 5 %) журналов с самым высоким фактором влияния (панели А на рис. 4 и 5). Наоборот, панели Б этих рисунков показывают долю верхних 10 % (и 5 %) большинства статей, которые не публикуются в верхних 10 % (и 5 %) журналов с наивысшим фактором влияния. Оба рисунка показывают, что взаимосвязь между фактором влияния и цитированиями постоянно ослабляется с 1990 г., так как бóльшая доля верхних (5 и 10 %) наиболее цитируемых статей публикуется вне журналов с верхним фактором влияния (5 и 10 %).

Весьма характерно, что доля 10 % наиболее цитируемых статей, опубликованных в 10 % самых цитируемых журналов, уменьшилась с 1990 г. (рис. 4, А) с почти 5,25% до 4,50%. Соответственно, доля 10% наиболее цитируемых статей, **не** опубликованных в 10% журналов с наивысшим фактором влияния, увеличилась с 1990 г. (рис. 4, Б) с 52% до почти 56%. Этот пример является, несомненно, более очевидным, когда те же самые сравнения делаются для верхних 5% статей и верхних 5% журналов (рис. 5). В 1990 г. примерно 2,25% верхних 5% статей были опубликованы в верхних 5% журналов; но к 2009 г. эта цифра снизилась до 1,90% (рис. 5, А). Подобным образом в 1990 г. около 55% верхних 5 % наиболее цитируемых статей **не** были опубликованы в верхних 5% журналов, но к 2009 г. данная цифра увеличилась до 62% (рис. 5, Б). Следовательно, наиболее важная литература в возрастающей степени приходит из большего круга журналов, а не только из журналов с наивысшим фактором влияния.

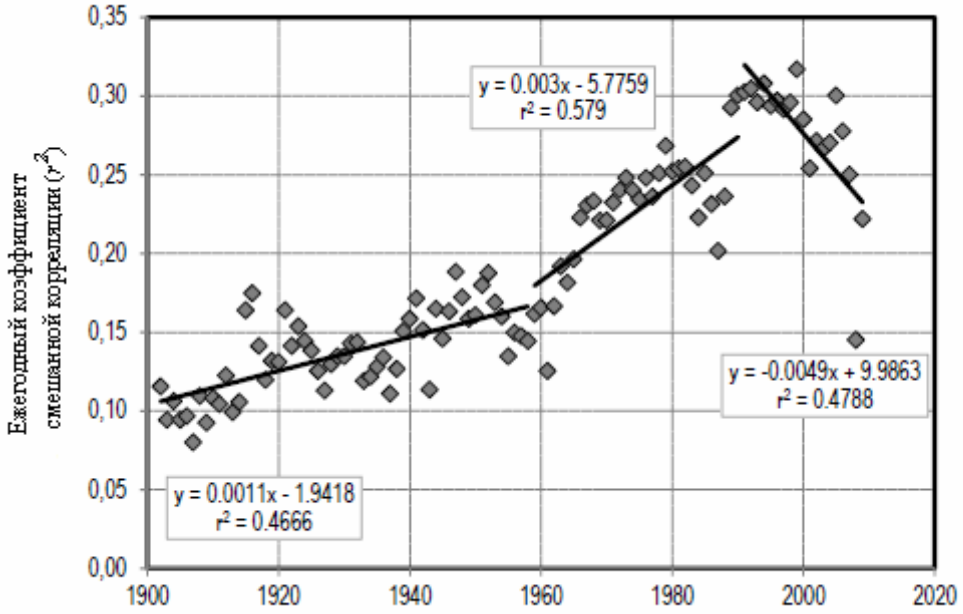


Рис. 1. Коэффициент смешанной корреляции (r^2) между фактором влияния журналов и двухгодичной частотой цитирования их статей за 1902 – 2009 гг. для всех журналов по естественным и медицинским наукам

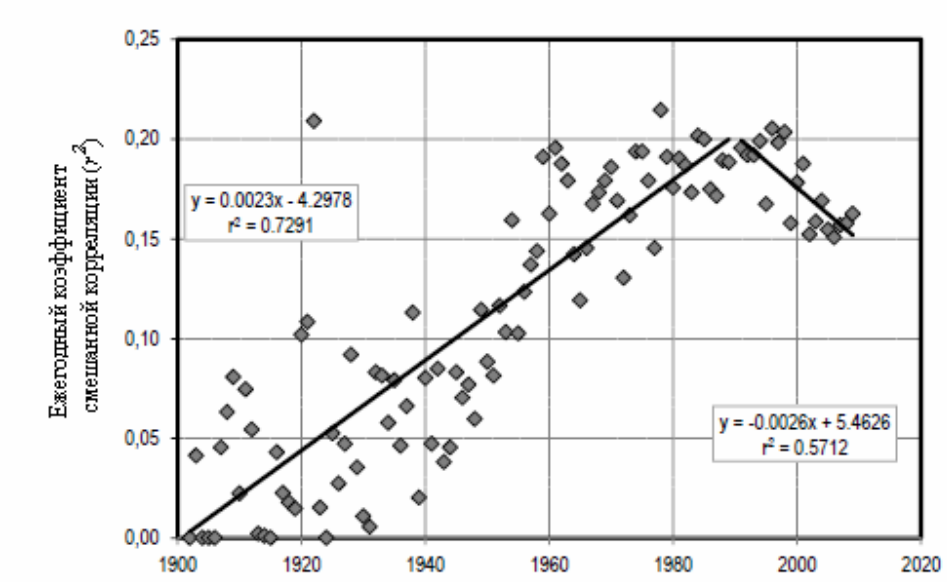


Рис. 2. Коэффициент смешанной корреляции (r^2) между фактором влияния журналов по физике и двухгодичной частотой цитирования их статей за 1902 – 2009 гг.

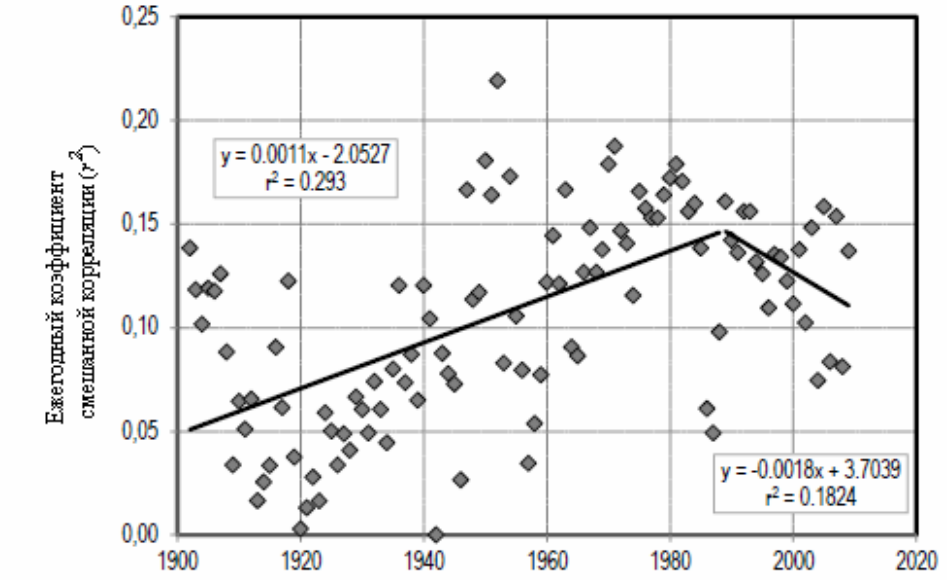


Рис. 3. Коэффициент смешанной корреляции (r^2) между фактором влияния журналов по общественным наукам и двухгодичной частотой цитирования их статей за 1902 – 2009 гг.

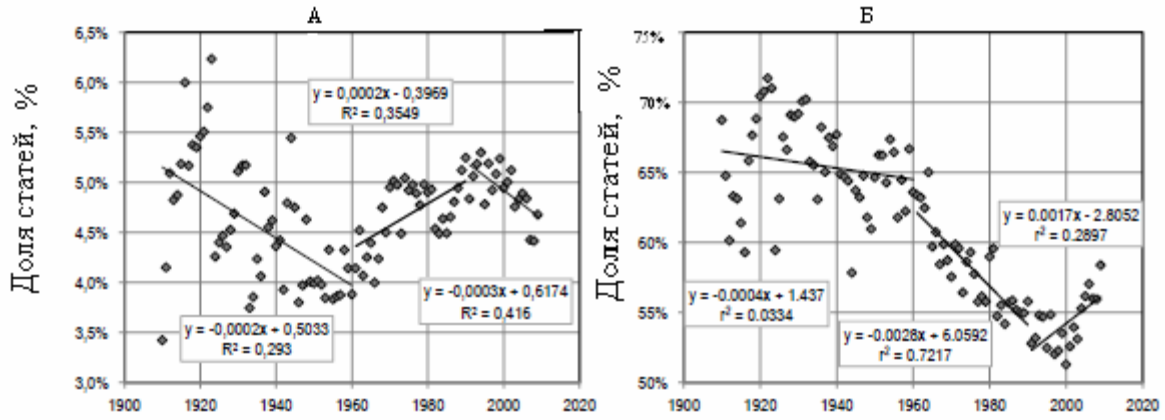


Рис. 4. А. Доля входящих в верхнюю часть 10 % наиболее цитируемых статей, опубликованных в верхних 10 % наиболее цитируемых журналов. Б.Доля верхних 10 % наиболее цитируемых статей, которые не публиковались в верхних 10 % наиболее цитируемых журналов.

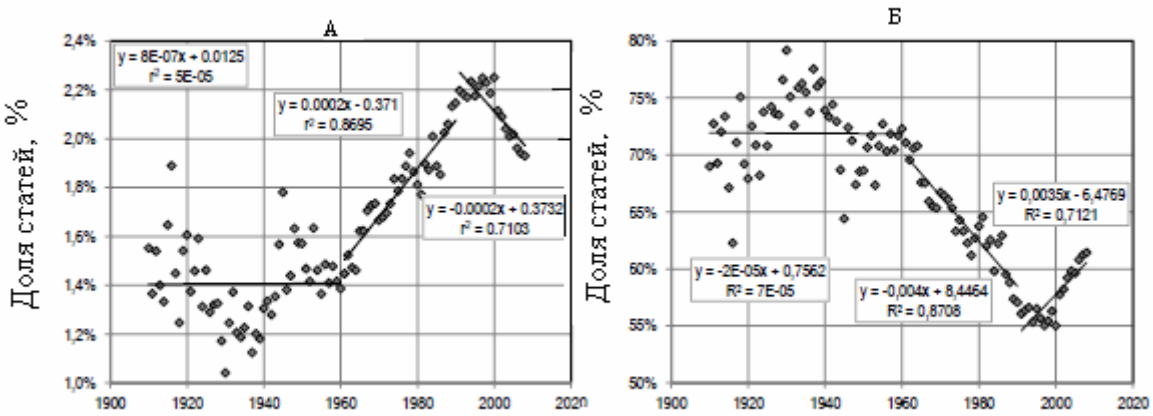


Рис. 5. А. Доля верхних 5 % наиболее цитируемых статей, опубликованных в верхних 5 % наиболее цитируемых журналов. Б. Доля верхних 5 % наиболее цитируемых статей, которые **не** были опубликованы в верхних 5 % наиболее цитируемых журналов.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ И ВЫВОДЫ

Факторы влияния первоначально разрабатывались для того, чтобы помочь библиотекам в их решениях относительно приобретения журналов и, следовательно, они оказывали важное влияние на обращение и доступность журналов. Здесь мы показываем, что на протяжении большей части XX в. связь между фактором влияния и цитированиями статей становилась сильнее, но, как предсказывалось, эта связь постоянно ослаблялась с наступлением электронного века. Кажется, что это изменение началось несколько раньше в физике, области, которая быстрее приняла электронное распространение информации. Также в это время доля статей верхнего ранга, пришедших из журналов верхнего ранга, снизилась. Наряду с тем фактом, что в целом цитирования стали более широко распространенными среди журналов [5], электронный век и его новые методы распространения и доступа к научной литературе могут привести к исчезновению фактора влияния как полезного критерия качества журналов, статей и ученых и иметь интересные последствия для будущего научной литературы.

Фактор влияния неоднократно критиковался с точки зрения подходящего критерия качества журналов [7– 13]. Самыми сильными аргументами против его значимости и использования являются: 1) некоторые виды публикаций в журналах, такие как письма и комментарии, используются для подсчета ссылок (номинатор), но сами не считаются «статьями» (деноминатор), и, следовательно, раздувают фактор влияния журнала; 2) фактор влияния зависит от числа ссылок, которое разнится среди дисциплин и журналов; 3) включение журналов в базу данных зависит только от издательства Thomson Reuters, частной компании, а не от сфер деятельности профессионалов; 4) точный фактор влияния, опубликованный издательством Thomson Reuters, не может быть повторен с использованием публично доступных данных; 5) распределение ссылки / статья не является обычным, так что, по крайней мере этот способ или медиана должны использоваться вместо среднего значения; 6) двухгодичный промежуток времени для статей с одним годом цитирования является полностью произвольным и благоприятствует высоко обращающимся, а не долго действующим статьям; 7) редакторы журналов могут манипулировать своими факторами влияния и искусственно их завышать. Наш анализ определяет еще одну проблему: взаимосвязь между качеством статьи и фактором влияния ослабевает, таким образом, фактор влияния теряет свое значение как критерий качества журнала.

Второе, факторы влияния используются как показатели качества статей. За исключением совсем недавно опубликованных статей, которые пока не имели шанса быть цитированными, не существует причины использовать фактор влияния как показатель качества статьи. Каждый может легко иметь доступ к частоте цитирования любой отдельной статьи и определять, как статья сама по себе оценивается, независимо от фактора влияния ее журнала. Поскольку взаимосвязь между частотой цитирования статей и фактором влияния продолжает ослабевать и поскольку более важные статьи все в большей степени появляются в самых разных местах, то даже менее оправдано автоматически переводить репутацию и символический капитал журнала также на его совсем недавно опубликованные статьи. Это должно потребовать возврата к непосредственным оценкам качества статей в результате реального их чтения.

Третье, и даже более беспокоющее, - это трехшаговый подход использования фактора влияния для заключения относительно качества журнала, распространения его также на статьи, а затем на использование его для оценки ученых. Наши данные показывают, что журналы с высоким фактором влияния теряют свой оплот в качестве единственных хранилищ высококачественных статей и поэтому не являются легитимной основой для распространения фактора влияния журнала на его статьи и, менее всего, на отдельных ученых. Это совпадает с открытием, что за последнее десятилетие в экономике доля статей в журналах верхнего ранга, подготовленных людьми из ведущих учреждений, снизилась [14]. Более того, учитывая, что ученые могут оцениваться с использованием множества других критериев и библиометрических показателей [например, 15 - 19], оценка ученых путем простого просмотра факторов влияния журналов, в которых они публикуются, является как наивной, так и неинформативной.

За последние несколько веков журналы были удобным способом организовывать статьи по предмету, но поисковые системы теперь позволяют нам находить отдельные статьи по определенным темам из полного спектра журналов, таким образом, высоко предметно-специфические журналы могут стать устаревшими или начать объединяться. Электронные журналы открытого доступа, такие как семейство журналов PLoS, и элек-

тронные базы данных, такие как система ArXiv и ее родственные системы, будут продолжать получать преимущество. Используя такие хранилища открытого доступа, эксперты могут найти публикации в своих соответствующих областях и решить, какие стоит читать и цитировать, независимо от журнала. Если эта модель продолжится, то фактор влияния перестанет быть значимым критерием качества журналов, статей и ученых.

Благодарность. Мы благодарим Жана-Пьерра Робитайя за комментарии к раннему предварительному варианту этой статьи. Жорж А. Лозано выражает благодарность Тартускому университету за предоставление ему свободного доступа к электронным фондам.

ЛИТЕРАТУРА

1. *Ellison G.* Is peer review in decline? // Economic Enquiry. – 2011. – Vol. 49, No. 3. – P. 635-657.

2. *Averch H.A* Exploring the cost-efficiency of basic research funding in chemistry // Research Policy. – 1989. – Vol. 18, No. 3. – P. 165-172.

3. *Leydesdorff L., Bornmann L.* Integrated impact indicators compared with impact factors: An alternative research design with policy implications // Journal of the American Society for Information Science and Technology. – 2011. – Vol. 62, No. 11. – P. 2133-2146.

4. *Lozano G.A* A new criterion for allocating research funds: "Impact per dollar" // Current Science. – 2010. – Vol. 99, No. 9. – P. 1187-1188.

5. *Lundberg R.* Lifting the crown-citation z-score // Journal of Informetrics. – 2007. – Vol. 1, No. 2. – P. 145-154.

6. *Põder E.* Let's correct that small mistake // Journal of the American Society for Information Science and Technology. – 2010. – Vol. 61, No. 12. – P. 2593-2594.

7. *Rossner M., Van Epps H., Hill E.* Show me the data // The Journal of Cell Biology. – 2007. – Vol. 179, No. 6. – P. 1091-1092.

8. *Aksnes D.W.* Characteristics of highly cited papers // Research Evaluation. - 2003. – Vol. 12, No. 3. – P. 159-170.

9. *Morgan V., Janca A.* Revisiting the journal impact factor // Australasian Psychiatry. – 2000. – Vol. 8, No. 3. – P. 230-235.

10. *Rothenberg R.* The impact factor follies // Epidemiology. – 2008. – Vol. 19, No. 3. – P. 372.

11. *Seglen P.O.* Why the impact factor of journals should not be used for evaluating research // British Medical Journal. – 1997. – Vol. 314, No. 7079. – P. 498-502.

12. *Vakil N.* The journal impact factor: judging a book by its cover // American Journql of Gastroenterology. – 2005. – Vol. 100, No. 11. – P. 2436-2437.

13. *Whitehouse G.H.* Citation rates and impact factors: Should they matter? // British Journal of Radiology. – 2001. – Vol. 74, No. 877. – P. 1-3.

14. *Ellison G.* Is peer review in decline? // Economic Enquiry. – 2011. – Vol. 49, No. 3. – P. 635-657.

15. *Averch H.A* Exploring the cost-efficiency of basic research funding in chemistry // Research Policy. – 1989. – Vol. 18, No. 3. – P. 165-172.

16. *Leydesdorff L., Bornmann L.* Integrated impact indicators compared with impact factors: An alternative research design with policy implications // Journal of the American Society for Information Science and Technology. – 2011. – Vol. 62, No. 11. – P. 2133-2146.

17. *Lozano G.A* A new criterion for allocating research funds: "Impact per dollar" // Current Science. – 2010. – Vol. 99, No. 9. – P. 1187-1188.

18. *Lundberg R.* Lifting the crown-citation z-score // Journal of Informetrics. – 2007. – Vol. 1, No. 2. – P. 145-154.

19. *Põder E.* Let's correct that small mistake // Journal of the American Society for Information Science and Technology. – 2010. – Vol. 61, No. 12. – P. 2593-2594.

Если вы не можете найти журнал, существует ли он? Доступность журналов по библиотековедению и информатике в Интернете*

Тове Фабер ФРАННСЕН (Tove Faber FRANDSEN)

Университет Южной Дании, г. Оденсе, Дания

Йеппе НИКОЛАИСЕН (Jeppe NICOLAISEN),

Королевская школа библиотековедения и информатики, г. Копенгаген, Дания

Публичный доступ к всемирной сети стал широко распространенным в конце 1990-х – начале 2000-х гг., и сегодня документы часто публикуются в Интернете. Признается, что открытый доступ к научной литературе постоянно растет, так как все больше производителей и издателей научной литературы делают свои публикации бесплатно доступными через Интернет. Предполагается, что можно утверждать, что только научные Интернет-документы, найденные с помощью поисковой системы Google Scholar, можно считать существующими. Степень охвата Google Scholar, таким образом, является важным вопросом. В статье сообщается о результатах изучения 159 журналов в области библиотековедения и информатики и их степени охвата в Google Scholar. Выяснилось, что журналы, имеющие много номеров в год, не считаются более легкими для поиска, чем журналы, выходящие меньшим числом номеров. Журналы, публикуемые не на английском языке, и журналы открытого доступа заметно имеют более низкую степень отыскиваемости. Эта тенденция проявляется даже ярче для журналов, которые являются журналами открытого доступа и издаются не на английском языке. Журналы открытого доступа и выходящие не на английском языке представляют весьма разнородные группы, и изменение в их степени отыскиваемости является более ощутимым в отличие от традиционных, доступных за плату журналов, которые больше похожи друг на друга в отношении их поиска.

ВВЕДЕНИЕ

Существует известная загадка в философии, представляющая собой что-то в таком роде:

Если дерево падает в лесу, и нет никого рядом, чтобы это услышать - создает ли дерево звук?

Эта загадка может восходить, по крайней мере, к Джорджу Беркли (1685-1753) и его работе «Трактат о началах человеческого знания». Задаваясь вопросами, в действительности, о наблюдении и реальности, эта загадка занимала многие светлые головы с тех пор. Основой загадки является, конечно, вопрос - может ли что-то существовать без того, чтобы не быть ощущаемым. Мы здесь не пытаемся решить загадку, а лишь бегло упоминаем ее, так как она похожа на такую же загадку, которую мы обнаружили в нашей собственной области – библиотековедении и информатике:

Если документ издан, но никто не может его найти, существует ли он?

Представьте, например, книгу, помещенную в несоответствующем месте в библиотечном фонде. Классификационный код в библиотечном каталоге говорит пользователю, где искать книгу, но там ее нет. Она ошибочно была помещена на другую книжную полку предыдущим пользователем. В таком случае существует ли книга? Подобно загадке о падающем в лесу дереве, эта загадка ставит вопрос – может ли что-то существовать без того, чтобы не быть ощутимым. Это метафизический вопрос, и ответ зависит отчасти, по крайней мере, от того, верите ли вы, что объект отличается от его свойств или объект является просто его чувственными данными.

Ставя вопрос о существовании на более конкретный уровень, ряд исследований показывает, что преподаватели и ученые стремятся больше приме-

* Перевод Frandsen T. F., Nicolaisen J. If you can't retrieve it, does it exist? Accessibility of LIS journals on the Internet//Procceedings of ELPUB 2011. Digital publishing and mobile technologies. 15th International conference on electronic publishing 22-24 June 2011, Istanbul, Turkey, pp. 87-94. — http://elpub.scix.net/cgi-bin/works/Show?114_elpub2011

нять легко доступные информационные источники, чем менее легко доступные информационные источники. Одним примером является исследование нейроученых, проведенное Вибертом и др. [1]. Это исследование предполагает, что когда ученые могут иметь доступ к эффективным и исчерпывающим электронным ресурсам BDI (BDI resources), эти ресурсы быстро становятся для них предподчительным способом получения связанной с работой информации. Это соответствует наблюдениям Тенопир и др. [2], которые обнаружили, что астрономы стремятся выбирать средства доступа, подходящие для печатного, электронного вида или обоих этих видов. Таким образом, предположение, выдвинутое Одлыжко и др. [3], о том, что ученые, издатели и библиотекари должны будут предпринимать даже большие усилия, чтобы сделать свои материалы легко доступными, является таким же справедливым сегодня, как и почти десять лет назад. В последние годы произошло главное изменение в способах, которыми научная коммуникация (формальная и неформальная) распространяется с помощью электронных средств. В то же время, поведение ученых при информационном поиске изменилось, особенно когда это касается использования поисковых систем и цифровых ресурсов, помимо традиционных журналов [4]. Электронная доступность журналов предполагает, помимо других вещей, что ученые сегодня реже посещают библиотеки. Просмотр сети и таблица сигнальных сообщений электронной почты о контентах постепенно вытесняет физический просмотр, а поиск становится популярной опцией, чтобы оставаться в курсе данных относительно новых достижений [5]. Однако, подобно неверно размещенной книге в библиотечном фонде, можно спросить, существует ли реально эта литература открытого доступа. Равноценно ли «размещение чего-либо в Интернете» претензии на существование? Мы так не считаем. Интернет огромен. Значит, размещение чего-то в Интернете не является само по себе претензией на существование. Оно подобно неправильно размещенной книге в библиотечном фонде. Нет никакой гарантии того, что публикация будет когда-нибудь найдена снова. Не вдаваясь в детали метафизического вопроса относительно существования, мы считаем, что первым требованием существования Интернет-документа является то, чтобы он был найденным. Это, конечно, ставит вопрос о том, что значит быть отыскиваемым в Интернете. Все документы в Интернете имеют уникальный URL (унифицированный указатель ресурсов). Таким образом, в принципе все документы в Интернете являются отыскиваемыми – надо только знать правильный URL. Снова это сравнимо с неправильно помещенной книгой в библиотечном фонде. Можно найти книгу, если знать, где смотреть. Таким образом, одного URL недостаточно. Ограничивая обсуждение научными документами, мы будем утверждать, что документ является отыскиваемым, если его можно найти путем поиска в подходящей поисковой системе. Какая тогда система является подходящей для поиска научных документов в Интернете? Мы считаем, что это Google Scholar (scholar.google.com). Это убеждение основано отчасти на ее предполагаемом предпочтении учеными [6] и отчасти на отчетах, утверждающих, что Google Scholar гарантирует высокую точность в поисковых результатах в сравнении с подобными поисковыми системами, благодаря ее более узкому определению научной литературы:

«Схожая в подходе, но более широкая и менее специфичная по тематике, чем Google Scholar, научная поисковая система Scirus (www.scirus.com) осуществляет поиск, который, согласно информации, обеспечивает примерно 300 млн специальных научных сетевых страниц. Помимо научных документов из Elsevier (сервер ScienceDirect, www.sciencedirect.com) предоставляются бесплатно доступные документы, многие из публичных сетевых серверов научных институтов. Среди них, например, имеются документы, размещенные студентами, которые не удовлетворяют научным критериям, таким как рецензирование, что часто приводит к их исключению из поиска. По нашему опыту в указателе Scirus существует совсем незначительная доля записей из неакадемических сетевых пространств Google Scholar. Охват системой Scirus чисто научных источников, помимо полнотекстового массива Elsevier's ScienceDirect, является низким (сравни выборку хостов в интерфейсе расширенного поиска в Scirus, http:scirus.com/ srsapp/advanced). То, что Scirus декларирует как «остальную научную сеть», является слишком общей, специально нефильтрованной, и компенсирует большинство выдач на любой запрос» [7, с. 815-816].

Значит, если нужно найти научные документы в Интернете, то Google Scholar является сегодня местом, куда стоит идти и искать*. Но насколько хорошо действует Google Scholar? В какой степени она охватывает научную литературу в Интернете?

Исследование, проведенное Ченом [13], обнаружило, что Google Scholar охватывает от 98 до 100% научных журналов как из публично доступного сетевого контента, так и из баз данных по подписке, с которыми сотрудничает Google Scholar. Левандовски [14] измеряет охват Google Scholar для совокупности журналов в области библиотековедения и информатики, как определено в списке ядерных журналов по библиотековедению и информатике в исследовании авторов [15]. Он проверил каждую статью из 35 основных журналов по библиотековедению и информатике за 2004-2006 гг. на доступность в Google Scholar. Информация о типе доступности (была ли доступна конкретная статья как файл PDF за плату, как файл PDF бесплатно или как препринт) была проанализирована и распределена по типу издательства. Однако исследование не включает открытый доступ как особую характеристику журналов при анализе. Левандовски [14, с. 257] утверждает:

«Цифры для препринтов несколько разочаровывают, так как на открытый доступ возлагаются боль-

* Мы признаем, что у Google Scholar также существуют недостатки и проблемы (см. [8-9] и http://en.wikipedia.org/wiki/Google_Scholar#Limitations_and_crititism), но по нашим сведениям ни одна другая поисковая система сегодня не является соответствующей. В одних областях Google Scholar действует лучше в терминах полноты и точности, чем большинство баз данных по подписке при представлении простых запросов из ключевых слов [10], тогда как в других областях Google Scholar соответствует по результатам базам данных по подписке в терминах полноты, а не точности [11-12].

шие надежды и имеется готовность авторов сделать свои работы доступными через открытый доступ. Особенно для специальности библиотековедения и информатики, с ее многими сторонниками открытого доступа, эти цифры кажутся весьма низкими».

Вывод Левандовски поддерживается Веем [16, с. 302], который заключает, что «архивирование статей не является регулярной практикой в данной области». Вей [16] измеряет доступность в открытом доступе исследований по библиотековедению и информатике в Google Scholar путем поиска статей из 20 ведущих журналов по библиотековедению и информатике. Это исследование не сосредоточено на журналах открытого доступа, но делает вывод, что Google Scholar является эффективным поисковым средством для поиска статей по библиотековедению и информатике.

Открытый доступ может быть также в форме журналов открытого доступа, и значит, что мы хотели бы рассмотреть вопрос об охвате Google Scholar в этом исследовании путем включения дополнительного числа журналов открытого доступа. Эта статья, конечно, является исследованием охвата поисковой системы Интернета, но, как мы пытаемся утверждать, наша тема затрагивает более глубокие философские вопросы.

МЕТОДЫ

Что касается этого исследования, то мы хотим изучить вопрос, является ли открыто доступная журнальная литература по библиотековедению и информатике отыскиваемой в Google Scholar независимо от формы индексируемой публикации (например, препринт или издательская версия). Прежде всего, нам необходимо сделать выборку из массива литературы по библиотековедению и информатике на основе некоторых формальных характеристик. Чтобы создать наш набор данных мы использовали список журналов по библиотековедению и информатике, доступный в *Journal Citation Reports (JCR)* и *Directory of Open Access Journals (DOAJ)*. Список первой базы данных содержит 66 журналов, а второй – 117. Исключая дубликаты, наш набор данных для этого исследования состоял из 177 журналов (см. табл. – список журналов). В этом исследовании учитываются все типы документов. Ряд публикаций, изданных в 2009 г., был проверен по сетевым сайтам журналов. Никаких отличий между различными типами публикаций не делалось. Очевидно, что некоторым типам публикаций исследование отдает большее предпочтение, чем другим. Только такие публикации, которые предусматривают инструкции для статей или авторов, были исключены из этого исследования. В разделе «Обсуждение результатов и выводы» мы вернемся к тому, как последующие исследования могут подтвердить этот набор данных.

Сейчас нам необходимо определить показатели, чтобы включить их в исследование.

Зависимый показатель есть число отыскиваемых документов в Google Scholar. Необходимо ввести понятия «видимость», «отыскиваемый» или «индексируемый в Google Scolar». Публикацию трудно назвать видимой или отыскиваемой, если она появляется в Google Scholar только с библиографической информацией в библиографии. Число точек предметного доступа относительно мало, и мы не получим доступа к полной публикации, реферату или предметным рубрикам. В этом исследовании документ определяется как отыскиваемым в Google Scholar, если этот документ индексируется с помощью точек предметного доступа помимо библиографического описания, т.е. с помощью реферата, полного текста или предметных рубрик. В некоторых случаях платный доступ мешает нам получить доступ к полному документу, хотя он представлен в Google Scholar. Каждый журнал проверялся в Google Scholar по 2009 году публикации. Норузи [17] суммирует поисковые методы, доступные через Google Scholar. Петер Яско [18,19] утверждает, что размер базы данных значительно увеличился за 2005-2008 гг., и Google Scholar прежде всего не хватает удовлетворительных уровней правильного назначения названий журналов и имен авторов, что спровоцировало трудности в поиске статей. Однако произвольное рассмотрение показывает, что эта проблема является нетипичной в этом случае. Здесь представленные запросы были основаны на названии журнала и годе публикации с использованием формуляра расширенного поиска в Google Scholar.

Более того, нам необходимо определить соответствующие независимые показатели. Формы открытого доступа могут быть независимыми переменными. Левандовски [14] обнаруживает, что доступность препринтных версий статей является в некотором роде неутешительной, так как только малое число статей доступно в этой форме. Автор обнаруживает, что это не связано с индексированием в Google Scholar, а скорее является результатом отсутствия публикаций, самоархирвируемых авторами в области библиотековедения и информатики. Это заставляет автора сделать вывод, что «хотя многие ученые и специалисты-практики в области библиотековедения и информатики являются сторонниками открытого доступа, но когда он сводится к расширенному открытому доступу через депонирование препринтов, их поддержка ограничивается» [14, с. 260]. Следовательно, мы предпочитаем не разделять доступность по формам. Мы сосредоточены на том, представлены ли эти статьи (независимо от того, имеется ли открытый доступ к их полнотекстовой версии).

Во-вторых, необходимо рассматривать язык как независимую переменную. Левандовски [14, 20] считает, что значительно высокий процент статей по библиотековедению и информатике на английском языке доступен в Google Scholar. Тем не менее, журналы, выходящие не на английском языке, обычно издаются небольшими издательствами, и у некоторых из них нет доступной электронной версии, поэтому существует более низкий процент охвата. Автор делает вывод: «эти результаты не показывают смещение по языку или стране в сторону статей на английском языке» [15, с. 260]. Мы хотели бы дополнить исследование журналов не на английском языке другими языками, отличными от немецкого языка, и журналами не на английском языке с электронными версиями. Значит, язык добавляется как показатель в наш набор данных.

Методы нашего исследования придерживались такой процедуры: каждый журнал входил в набор

данных с ISSN, чтобы однозначно определить журнал, название журнала, замена для языка (неанглийский язык=1, если журнал включал публикации на других, нежели английский, языках, в противном случае 0), замена для открытого доступа (OA=1 для открытого доступа, иначе 0), число публикаций в 2009 г. (в соответствии с сетевыми сайтами журналов или альтернативно указателям ссылок) и число публикаций, проиндексированных в Google Scholar. На основе этих данных мы можем подсчитать долю публикаций, проиндексированных в Google Scholar.

РЕЗУЛЬТАТЫ

Прежде всего, нам необходимо рассмотреть возможную корреляцию между числом изданных публикаций и их долей, доступной через Google Scholar. Можно утверждать, что небольшим журналам (т.е. изданиям с малым числом ежегодных выпусков) трудно привлечь внимание вообще и особенно Google Scholar, тогда как крупные журналы (т.е. издания с большим числом выпусков в год) являются, как правило, более заметными. На рис. 1 иллюстрируется соотношение между числом публикаций и их долей, доступной в Google Scholar. Один выброс (резко выделяющееся соотношение) был исключен из рис.1, так как он значительно нарушал картину.

Из рис. 1 ясно, что такая корреляция не может быть обнаружена. Статистические анализы подтверждают картину, изображенную на рисунке. Линейные регрессии (с или без выброса) с долей в Google Scholar в качестве зависимой переменной и публикаций как независимой переменной открывают *p*-значения независимой переменной, ранжируясь от 0,25 и 0,29. Далее мы переходим к анализам.

В таблице дается обзор средних и медианных значений. Средние значения обычно используются для описания основной тенденции набора данных, которые не имеют экстремальных значений (выбросов). С другой стороны, медиана обычно используется для описания основной тенденции набора данных, которые имеют экстремальные значения, так как она не подвержена влиянию выбросов настолько сильно, как средние значения. В таком случае значения ранжируются от 0 до 1, и выбросов нет (см. рис.1и таблицу). Следовательно, среднее значение является подходящей мерой для описания середины.

В следующих рисунках мы используем среднее значение, чтобы проанализировать различия в охвате. На рис. 2 иллюстрируется обозримость в Google Scholar журналов открытого доступа и традиционных доступных за плату журналов. Планки погрешностей представляют 95 % доверительного интервала. Мы можем ясно видеть, что хотя имеется некоторое отличие среди журналов внутри двух категорий, традиционные доступные за плату журналы имеют бóльшие проценты охвата в Google Scholar, чем журналы открытого доступа. Первые имеют средний процент охвата в 79%, тогда как вторые - 52%.

На рис. 3 отражается обозримость в Google Scholar журналов, публикуемых исключительно на английском языке, и журналов, публикуемых как на английском, так и на других языках. Журналы, публикуемые исключительно на английском языке, имеют более высокую обозримость в Google Scholar, так как их доля в Google Scholar на 17 % выше, чем тех, что публикуются как на английском, так и на других языках. Последняя группа, кажется, является более разнородной группой журналов, так как их вариативность больше.

На рис.4 отражается обозримость журналов открытого доступа в Google Scholar, публикуемых как на английском, так и на других языках, по сравнению с доступными за плату журналами, публикуемыми исключительно на английском языке.

Журналы открытого доступа, включающие публикации не на английском языке, характеризуются даже более низкой долей обозримости в Google Scholar. Оставшиеся журналы набора данных не могут быть охарактеризованы в терминах средних значений иначе, чем неанглийские журналы. Тем не менее, их доля на 11% ниже, чем доля доступных за плату журналов.

Обзор результатов

Доля публикаций в Google Scholar							
		Число публикаций N	Среднее значение	Медианное значение	Процентиль 05	Процентиль 25	Процентиль 75
Открытый доступ	0	62	0,79	1,00	0,10	0,63	1,00
	1	97	0,52	0,50	0,00	0,04	1,00
Неанглийский язык	0	108	0,68	0,88	0,00	0,40	1,00
	1	51	0,51	0,26	0,00	0,00	1,00
Открытый доступ для неанглийского языка	0	114	0,68	0,88	0,00	0,36	1,00
	1	45	0,50	0,23	0,00	0,00	1,00

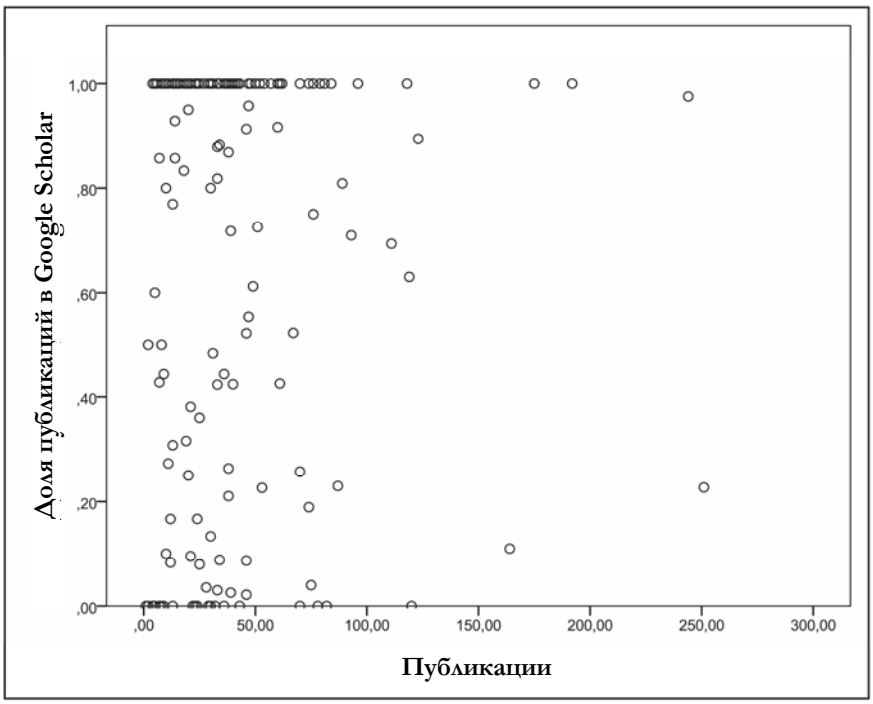


Рис. 1. Доля публикаций в Google Scholar и число выпусков, публикуемых ежегодно журналом

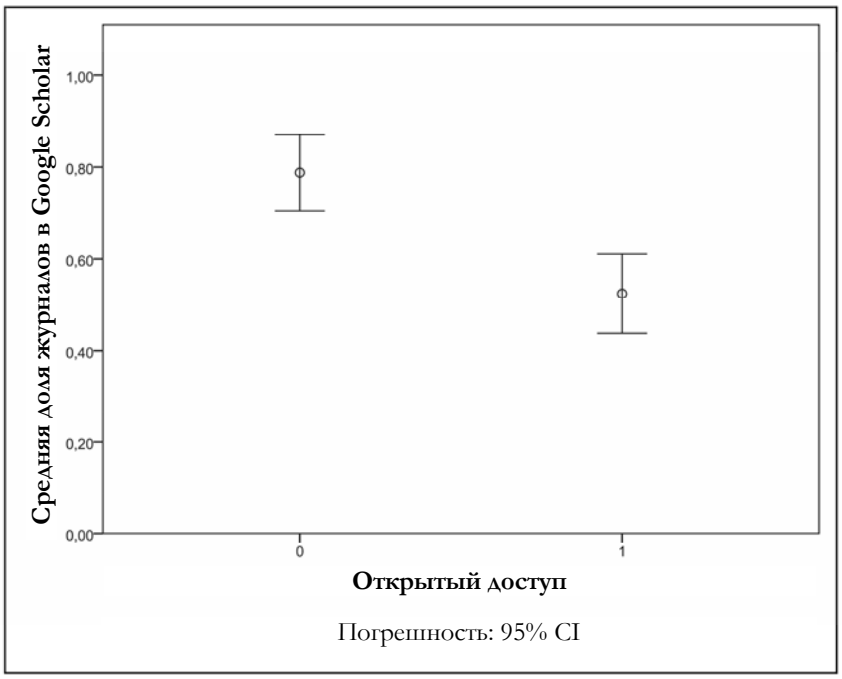


Рис. 2. Средняя доля проиндексированных в Google Scholar журналов и статус открытого доступа. Планки погрешностей представляют 95% доверительного интервала (CI)

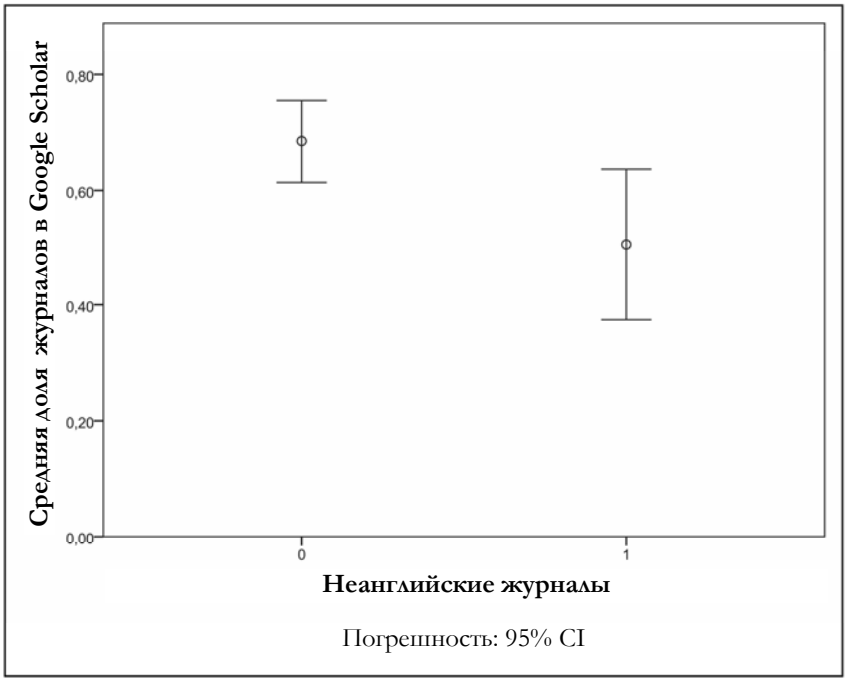


Рис. 3. Средняя доля проиндексированных в Google Scholar журналов и язык. Планка погрешностей: 95% доверительного интервала (CI)

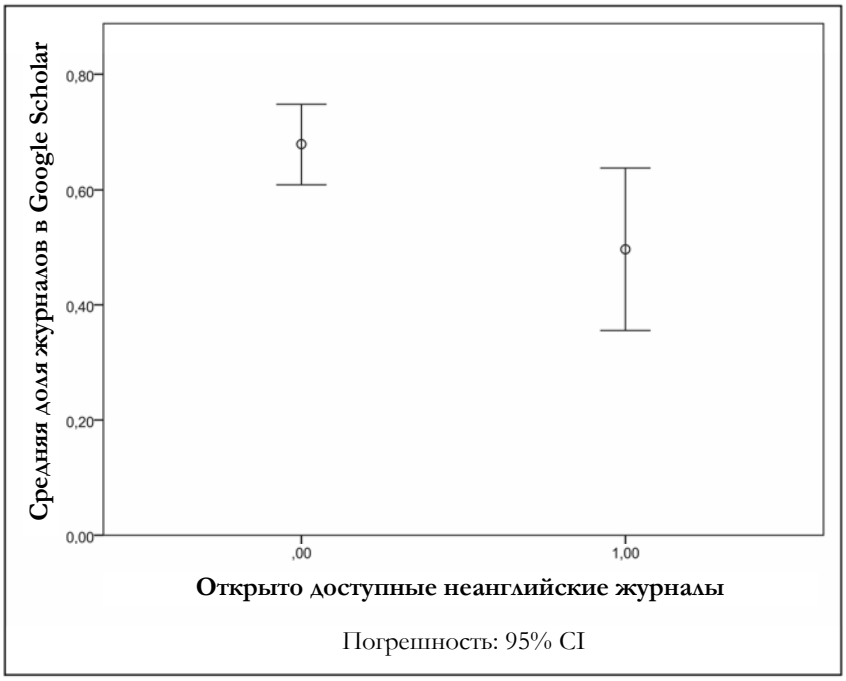


Рис.4. Средняя доля проиндексированных в Google Scholar журналов и журналов открытого доступа, а также публикуемых не на английском языке. Планка погрешностей представляет 95% доверительного интервала (CI)

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ И ВЫВОДЫ

До обсуждения применения наших результатов мы кратко суммируем наши наблюдения. Прежде всего, крупные журналы (т.е. журналы с большим числом ежегодных выпусков) не являются в большей степени отыскиваемыми в Google Scholar, чем небольшие журналы. Журналы не на английском языке и журналы открытого доступа имеют склонность к более низкой степени отыскиваемости в Google Scholar. Эта тенденция является даже более сильной для журналов как открытого доступа, так и издаваемых не на английском языке. Кроме того, журналы открытого доступа и журналы не на английском языке являются весьма разнородными группами, и существует гораздо большая вариативность в степени их отыскиваемости, чем в случае традиционных, доступных за плату журналов, которые схожи между собой в отношении отыскиваемости.

Как мы пытались прояснить на протяжении статьи, открытый доступ также является вопросом отыскиваемости, но не должно быть корреляции между открытым доступом и отыскиваемостью. Размещение чего-либо в Интернете бесплатно является щедрым, но если никто не может определить местонахождения этого документа и получить доступ к нему, едва ли это дает различие.

Настоящая работа может быть расширена и обогощена включением большего числа журналов в исследование. Настоящее исследование вносит ясность в отношение между открытым доступом и отыскиваемостью для открытого доступа и журналов с достаточным международным влиянием, чтобы быть включенными в JCR. Ряд журналов с иными характеристиками, чем журналы в этом исследовании, такие как доступные за плату журналы из развивающихся стран, могли бы помочь нарисовать более богатую картину связи между открытым доступом и отыскиваемостью.

Используя Google Scholar в качестве нашего показателя отыскиваемости, мы обнаружили большие различия в отыскиваемости среди журналов открытого доступа. Ответственность за это лежит на редакторах. Кажется, что некоторые журналы открытого доступа могли бы выиграть от более активного подхода к тому, чтобы иметь их контент, представленным в Google Scholar. Имеются также технические руководства включения, а также форма запроса включения, доступная на сетевом сайте Google Scholar (http://scholar.google.com/intl/en/scholar/publishers.html#faq1). Это исследование подтверждает, что редакторам журналов открытого доступа следует осознать важность соблюдения этих руководств, так как в известном смысле: *Если вы не можете чего-либо найти, то этого не существует!*

ЛИТЕРАТУРА

1. *Vibert N., Rouet J-F., Ros C., Ramond M., Deshoullieres B.* The use of online electronic information resources in scientific research: The case of neuroscience// Library & information science research. — 2007. — Vol. 29, No.4. — P. 508 - 532.

2. *Tenopir C., King D.W., Boyce P., Grayson M., Paulson K-L.* Relying on electronic journals: Reading patterns of astronomers// Journal of the American Society for Information Science and Technology. — 2005. — Vol.56, No.8. — P. 786 - 802.

3. *Odlyzko A.* The rapid evolution of scholarly communication// Learned Publishing. — 2002. — Vol. 15, No.1. — P. 7-19.

4. *Meyer E.T., Schroeder R.* The world wide web of research and access to knowledge// Knowledge Management Research & Practice. — 2009. — Vol.7, No.3. — P. 218-233.

5. *Olle C., Borrego A.* A qualitative study of the impact of electronic journals on scholarly information behavior//Library and Information Science Research. — 2010. — Vol. 32, No. 3. —P. 221-228.

6. *Jamali H.R., Asadi S.* Google and the scholar: The role of Google in scientists' information-seeking behavior// Online Information Review. — 2010. — Vol.34, No.2. — P. 282-294.

7. *Mayr P., Walter A.-K.* An exploratory study of Google Scholar// Online Information Review. — 2007. — P. 31, No.6. — P. 814- 830.

8. *Jacso P.* Google Scholar's ghost authors// Library Journal. — 2009. — Vol. 134, No.18. — P. 26-27.

9. *Jacso P.* Metadata mega mess in Google Scholar// Online Information Review. — 2010. — Vol. 34, No.1. — P. 175-191.

10. *Walters W. H.* Google Scholar search performance: Comparative recall and precision// Portal-Libraries and the Academy. — 2009. — Vol. 9, No.1. — P. 5-24.

11. *Anders M.E., Evans D.P.* Comparison of PubMed and Google Scholar Literature Searches// Respiratory Care. — 2010. — Vol.55, No. 5. — P. 578-583.

12. *Freeman M.K., Lauderdale S.A., Kendrach M.G., Woolley T.W.* Google Scholar versus PubMed in locating primary literature to answer drug-related questions// Annals of Pharmacotherapy. — 2009. — Vol. 43, No.3. — P. 478-484.

13. *Chen X.T.* Google Scholar's dramatic coverage improvement five years after debut// Serials Review. — 2010. — Vol. 36, No.4. — P. 221-226.

14. *Lewandowski D.* Google Scholar as a tool for discovering journal articles in library and information science// Online Information Review. — 2010. — Vol. 34, No.2. — P. 250-262.

15. *Schloegl C., Petschnig W.* Library and information science journals: An editor survey//Library Collections, Acquisitions, & Technical Services. — 2005. — Vol. 29, No. 1. — P. 4-32.

16. *Way D.* The open access availability of library and information science literature// College & Research Libraries. — 2010. — Vol. 71, No.4. — P. 302-309.

17. *Noruzi A.* Google Scholar: The new generation of citation indexes// Libri. — 2005. — Vol. 55, No.4. — P. 170-180.

18. *Jacso P.* Google Scholar: The pros and cons// Online Information Review. — 2005. — Vol. 29, No.2. — P. 208-214.

19. *Jacso P.* Google Scholar revisited// Online Information Review. — 2008.—Vol. 32, No.1. — P. 102-114.

20. *Lewandowski D.* Nachweis deutschsprachiger bibliotheks- und informationswissenschaftlicher Aufsätze in Google Scholar// Information Wissenschaft und Praxis. — 2007. — Vol. 58, No.3. — P. 165-168.

Программа Freemium как поддерживающая экономическая модель для электронного издательства открытого доступа в области гуманитарных и общественных наук*

Пьер МУНЬЕР
(Pierre MOUNIER)

Центр открытого электронного издательства, EHESS/Cleo (Группа Cleo Школы высшего образования по общественным наукам), г. Париж, Франция

Из двух путей открытого доступа – зеленого и золотого – последний является более трудным для развития и имеет меньшую поддержку со стороны научного сообщества. Основная трудность состоит в нахождении надежной экономической модели. Обычно журналы открытого доступа зависят от двух источников финансирования: субсидии и/или денежные взносы учреждений и плата за публикацию от научных организаций в модели с авторской платой. Эти два пути финансирования журналов и книг открытого доступа оказались эффективными в ряде случаев (Plos), но не являются безупречными. Центр открытого электронного издательства, французская инициатива для издательства открытого доступа в области гуманитарных и общественных наук, недавно разработал новую экономическую модель, основанную на программе Freemium для своих журналов и книжных серий с полным открытым доступом, чтобы рассмотреть две проблемы: улучшить их экономическое положение и дать им бóльшую обозримость в библиотеках. "Freemium" - слияние слов "free" (бесплатный) и "premium" (премия, поощрение), сохраняет открытый доступ к информации наряду с маркетингом бесплатных услуг.

ВВЕДЕНИЕ

Движение за открытый доступ завоевало ряд побед с начала своего существования. Первая состояла в том, чтобы упрочить себя в качестве испытанного и легитимного метода передачи научных результатов. Этот успех является наиболее значительным, поскольку был достигнут посредством двух каналов: зеленая дорога, относящаяся к открытым архивам, и золотая дорога, касающаяся научных публикаций открытого доступа. В сфере открытых архивов ArXiv является самым поразительным успешным примером с 745 тыс. статей, депонированных самими авторами. В сфере публикаций открытого доступа – это платформа Plos со своими 7 журналами, которые являются наиболее часто цитируемыми, и это, таким образом, вполне справедливо. Тот факт, что растущее число коммерческих издателей, таких как Elsevier или Springer, предлагают круг публикаций открытого доступа, хотя они и не поддерживали упомянутое движение в течение ряда лет, является дальнейшим доказательством успеха. Было бы уместно для сторонников открытого доступа применить любимую фразу Ганди: «Сначала они игнорируют вас, затем они смеются над вами, потом они сражаются с вами, а потом вы побеждаете». Примеры успеха не должны мешать нам задавать вопросы относительно способа, каким действуют ресурсы открытого доступа в разных странах и дисциплинах. В гуманитарных науках и общественных науках, в частности, где в отличие от большинства научных дисциплин книга является более важным типом публикации, чем журнальная статья, положение совсем не такое простое.

Во Франции, например, большинство издателей в области гуманитарных наук противятся открытому доступу для своих журналов и книжных фондов. Оценку состояния достижений открытого доступа в стране сравнительно легко представить благодаря централизации платформ, публикующих электронный контент. Три основные французские платформы, связанные с публикациями в области гуманитарных и общественных наук, - это Carin.info, Revues.org и Persee.fr. Эта тройка публикует более 630 журналов, всего 62% общего числа действующих журналов в сфере дисциплин гуманитар-

* Перевод Mounier P. Freemium as a sustainable economic model for open access electronic publishing in humanities and social sciences. – http://elab.scix.net/cgi-bin/works/show?107_elpub2012

ных и общественных наук*. Из этого числа только 184 журнала издаются для полностью открытого доступа. Другие журналы ограничивают доступ к статьям через период эмбарго от одного года до пяти лет в зависимости от журнала. В то время как общее число статей, доступных в открытом доступе, может составлять более 400 тыс. статей, из-за инициатив цифрования ретроспективных фондов открытого доступа, таких как Persee**, число новых статей, публикуемых в открытом доступе каждый год, составляет только 4 тыс. из общего числа 22 тыс. статей, что является очень низкой долей.

В области гуманитарных и общественных наук книги составляют преобладающий и самый признанный способ передачи научных результатов. Вопрос открытого доступа для электронного распространения работ в области гуманитарных и общественных наук не являлся ключевой проблемой для книг до недавнего времени просто потому, что большинство из них не распространялось электронным способом. Однако в последние несколько лет благодаря ряду общих факторов, среди которых массовое цифрование книг через программу Google Books, которая, несомненно, является самой значимой, издатели начали переводить в цифровую форму свой каталог, хотя также распространяют новые публикации через электронные книжные хранилища. В преобладающем большинстве случаев работы, которые не являются частью публичной сферы, не доступны в открытом доступе. Инициативы, предпринятые новыми цифровыми издателями, такими как Open Humanities Press или Open Book Publishers, с одной стороны, и университетскими издательствами в рамках консорциумов, подобных Oapen, - с другой стороны, очень интересны с этой точки зрения. Однако они сравнительно маргинальны в отношении к продукции основных издателей в разных странах, таких как, например, Oxford University Press, Cambridge University Press, Presses Universitaires de France или De Gruyter, которые все предлагают ограниченный доступ.

Данная статья определяет два фактора, которые ставят под сомнение распространение публикаций открытого доступа в области гуманитарных и общественных наук: с одной стороны, экономические модели, в настоящее время используемые для этого способа распространения, являются слабыми, тогда как, с другой стороны, университетские библиотеки борются за сохранение своей традиционной роли документальных посредников контента открытого доступа. Статья также описывает недавний эксперимент, осуществленный в феврале 2011 г. французской платформой OpenEdition.org., известной как программа OpenEdition Freemium. Эта коммерческая модель, предназначенная для управления библиотеками с открытым доступом к контенту в области гуманитарных и общественных наук, нацелена на обеспечение прагматического и политического ответа на два препятствия, мешающие развитию открытого доступа, определенного ранее.

Экономические модели для публикации с открытым доступом

Существуют две экономические модели, которые преобладают в области открытого доступа к научным публикациям, среди них определенная Дж. Виллински как «десять особенностей открытого доступа» [1]. Одна из них основана на субсидиях, а другая представляет собой модель авторской платы.

Субсидии

Модель с субсидиями широко распространена среди публикаций в области гуманитарных и общественных наук, где издательство открытого доступа часто финансируется заранее с помощью грантов из различных научных учреждений (университеты, научные общества, фонды и правительства). Поддержка публикации принимает разные формы: финансовые, естественно, а также через обеспечение штатного персонала (редакторы рукописей статей и редакторы форматов, верстки) и рабочего пространства, хостинга области и технической поддержки. В случае Франции, например, журналы имеют бесплатный доступ к услугам платформ Revues.org и Persee.fr., которые были разработаны благодаря финансированию со стороны Министерства науки и высшего образования, а также таких организаций, как Национальный центр научных исследований (CNRS), Школа высшего образования по общественным наукам (EHESS), университеты Прованса, Авиньона и Лионский университет. Платформа Oapen была разработана благодаря финансированию со стороны европейских комиссий, а Scielo поддерживается рядом региональных, национальных и межправительственных агентств в Южной Америке. В то время как «инфраструктурная» поддержка в форме платформ распространения часто приходит от правительств или консорциумов в соответствии с их размером, поддержка ежедневной редакционной деятельности журналов в основном идет от локальных структур. Ряд журналов отражает деятельность определенного университетского отделения или научного общества. Перечень издателей журналов на платформе Revues.org демонстрирует это инвестирование со стороны университетских структур в распространении контента открытого доступа. Кроме коммерческих издательств и университетских издательств, существует множество университетских отделений, научных лабораторий и научных обществ.

Органически зависимая связь журнала с локальной научной структурой нуждается в тщательном изучении для выяснения вопроса, идеально ли она совместима с открытостью, необходимой для истинно научного журнала. Имеют ли журналы, основанные университетскими отделениями, которые часто полностью их финансируют, запас времени, чтобы избавиться от того, что называется «университетским локализмом (узостью)», мнение, которое многие считают вредным для качества публикации? Некоторые журналы, действительно, находят способы преодоления таких ограничений, чтобы через разнообразие своих авторов и их открытость стать базовыми журналами в их области исследования. Однако в таких ситуациях отделения университетов считают более трудным мобилизовать средства для поддержки журнала, который больше не является отражением работы отделения. Возможность получения коммерческого дохода от продажи выпусков и статей или от библиотечных подписок впоследствии часто ищется редакциями журналов не только в качестве основы для их эконо-

* По данным работы Archambault E., Vignola=Gargné E. L'utilisation de la bibliométrie dans les sciences socials et les humanités (2004), общее число французских журналов в области гуманитарных и общественных наук составляло примерно 2 тыс. журналов в 2004 г. В 2010 г. Институт научной и технической информации (INIST) создал перечень 1058 действующих во Франции журналов: http://www.inist.fr/spip.php?article85.

** Оценка основана на числе документов, представленных во французской поисковой системе Isidore: http://www.rechercheisidore.fr.

мической модели, но и как средства, чтобы избежать «собирания повсюду грантов» и получить научную и интеллектуальную независимость в отношениях со своими главными учреждениями.

Экономическая слабость, обусловленная моделью финансирования, полностью основанного на грантах, также должна приниматься во внимание, поскольку обеспечение технических средств, персонала, ссуд и грантов может быть прекращено в любой момент. Все это приобретает обратные тенденции, изменения в направлении или, как это происходит сейчас во многих странах, решительные финансовые ограничения означают, что щедрое финансирование может исчезнуть в любой день. В конечном счете, финансирование публикаций открытого доступа только с помощью субсидий оказывается очень слабой моделью с научной и экономической точек зрения, особенно когда это касается поддержки основы редакционной деятельности (рецензирования, редактирования рукописей), и не обеспечивает технический способ, которой должен быть электронным распространением через распределенные платформы.

Модель авторской платы

Другой доминирующей экономической моделью в сфере открытого доступа научных публикаций является модель «авторской платы». Подобно предыдущей модели, финансирование поступает до, а не после публикации, и состоит в основном в просьбе к авторам книг или статей оплатить издательские расходы по изданию их работ. Поскольку расходы оплачиваются автором, от читателей не просят денежных взносов, и, следовательно, контент может распространяться через открытый доступ. Это и есть способ, который был успешно использован изданием Public Library of Science (Plos) и обеспечил основы для его экономической жизнеспособности с момента создания, привлекая научные публикации высокого уровня. С 36 тыс. статей, представленных к публикации, и 17 тыс. статей, опубликованных в 2011 г., семейство журналов Plos (и особенно Plos One, его ведущий журнал) может быть справедливо названо «мегажурналом», динамики которого впечатляют. Стоимость публикации статей в журнале Plos колеблется от 1300 до 2900 долл. В среднем в 2010 г. было инвестировано 12 млн долл. непосредственно в Plos научными учреждениями, которые брали плату на публикационные затраты со своих ученых*.

Модель, которую предлагает Plos, стала настолько авторитетной в научной сфере, что она заставила других действующих участников, особенно коммерческих издателей, пересмотреть свою роль. Поэтому теперь коммерческие издатели включают в свою классическую коммерческую модель, основанную на ограниченном подпиской доступе или закупке, формулу «платит автор», позволяющую авторам сделать «бесплатными» свои публикации (статьи или книги), оплачивая публикационные расходы. Примерами служат издательство Shpringer с его формулой открытого выбора (Open Choice formula), корпорация Elsevier с ее программой спонсорства и издательство Oxford University Press с его открытой моделью (Oxford Open model). Аргументация, наиболее часто выдвигаемая сторонниками открытого доступа относительно поддержки этой модели, основана на совокупности затрат, связанных с исследованием и публикацией результатов. Например, Моррис [2] утверждает, что публикационные затраты составляют менее 20 % общих затрат на исследование, проведенное с целью создать содержание публикации. С точки зрения общей перспективы, на самом деле, кажется абсурдным для научных организаций, которые финансировали львиную долю исследовательской программы, впоследствии не добавлять дополнительную стоимость к маргинальным затратам на публикацию, включенную в распространение открытым доступом, что радикально улучшает влияние исследования, как доказал ряд авторов.

Однако следует быть осторожными в отношении глобальных перспектив. Несмотря на то, что модель «платит автор» может казаться подходящей для определенных ситуаций и дисциплин, вряд ли она будет широко принята другими. Вероятно, не случайно, что она так широко распространена в науках о жизни. В этих дисциплинах публикации в основном являются публикациями, подготовленными группой лиц, – статьи подписаны рядом авторов. В общественных науках справедливо обратное, огромное большинство статей подписано одним или, может быть, двумя авторами. Само написание является гораздо менее формальным и обусловленным; стиль имеет бóльшую свободу, чем в науках о жизни, и лучше отражает авторскую индивидуальность и оригинальность подхода, демонстрируя различную взаимосвязь между исследованием и публикацией в этих дисциплинах. Нельзя сказать, что авторы не должны иметь права на субсидии в области гуманитарных и общественных наук. Иногда это имеет место, и крупные издатели сферы гуманитарных и общественных наук реализуют системы открытого доступа, как и другие издатели в других дисциплинах. Однако трудно представить, что эта модель станет широко распространенной в области гуманитарных и общественных наук, если она не предполагает эффектов искажения или даже радикального выравнивания с преобладающими моделями публикации и способами написания материалов в экспериментальных науках.

Модель авторской платы также имеет еще одну более общую проблему - проблему самого определения открытого доступа. Одним важным аргументом в пользу открытого доступа является необходимость в преодолении финансовых барьеров к доступности информации, чтобы сделать научную коммуникацию более эффективной, поддержать финансовую рациональность и привести в действие этические соображения (особенно в науках о жизни). Тот же самый вопрос можно задать относительно системы допубликационного финансирования и особенно системы авторской платы. Если эта система станет более широко распространенной, то это, вероятно, создаст легко постижимые искажения в области исследования. Она обеспечивает механическое преимущество для научных групп в богатых исследовательских учреждениях (эффект Матфея). Она также оказывает экономическое давление на научные учреждения, тем же образом, как и классическая издательская система оказывает давление на библиотеки учреждений, потому что научный публикационный рынок является таким же негибким для ученых, которые публикуются, как и для подписывающихся библиотек. Существует еще одно слабое место, которое система авторской платы не учитывает: роль посреднических организаций в рамках документальной экосистемы открытого доступа, особенно роль, которую играют библиотеки.

* Все цифры взяты из работы Binfield Peter "The public library of science", представленной на 9-й Берлинской конференции 8 ноября 2011г., Washington DC, http://www.berlin9.org/bm-doc/berlin9-binfield.pdf.

МЕСТО БИБЛИОТЕК В ЗОЛОТОЙ ДОРОГЕ К ОТКРЫТОМУ ДОСТУПУ

«Особый характер модели открытого доступа состоит в том, что она раскрывает проблему: открытый доступ не влечет за собой подписку или приобретение. Все зависит от того, вовлечено ли учреждение в данное движение, одобрило ли оно членство для своих ученых с тем, чтобы они могли свободно публиковаться в журналах открытого доступа. Понятие подписки является существенным для других услуг и зависит от бюджета библиотек, тогда как членство финансируется из другого бюджета научного управления. Здесь нам надо быть осторожными. Грубо говоря, членство стоит денег, поэтому я возьму их из библиотечного бюджета, так как библиотека перестала платить за подписку, поэтому бюджет может быть перераспределен. Лично я никогда не слышал о таких грубых примерах, но то же объяснение может действовать безоговорочно … Что касается меня, то я склонен думать, что мы имеем посредническую роль: мы здесь для того, чтобы создать фонды и научить, как получить к ним доступ, используя наши услуги, но с точки зрения возможностей это не означает то же самое. Справедливо, что в настоящее время мы имеем большие бюджеты, поскольку ресурсы слишком дороги для приобретения. Наша роль определенно изменится, если изменится экономическая модель». Это выдержка из опроса по изучению Эммой Бестер в 2009 г. более двадцати научных библиотек, направленному на проблему обозримости ценного качества открытого доступа в секторах гуманитарных и общественных наук их учреждений [3]. Эта выдержка прекрасно суммирует основные результаты опроса. Его автор сочетает несколько методов, как количественных (оценка журналов регистрации сервера, опросы в реальном времени с замкнутыми анкетами), так и качественных (слабоструктурированные интервью, обозрение систем информационной технологии в библиотеках), чтобы оценить, в какой степени ресурсы открытого доступа, такие как предлагаемые Revues.org, цитировались, индексировались и рекламировались научными библиотеками. Результаты опроса поразительны и выдвигают одну особую проблему, которая редко упоминается. Когда библиотечный персонал, как правило, расположен в пользу открытого доступа на индивидуальной основе, когда персонал всякий раз, когда возможно, активно старается продвигать ресурсы открытого доступа, это обязательство противоречит тому способу, которым сами библиотеки действуют в роли документальных посредников, так как она мотивируется финансовым стимулом библиотек к приобретению документов. Интервьюируемый библиотекарь правильно оценивает ситуацию. Он объясняет традиционную роль библиотек: прежде всего, закупать ресурсы, а затем предлагать их пользователям. Таким образом, реальная проблема ресурсов открытого доступа в библиотеках является простой: они не могут быть закуплены. И, следовательно, библиотекарям трудно получить такие ресурсы и найти время, персонал и усилие для их эксплуатации, если не оказывается никакого преднамеренного, осознанного сопротивления. Причина проста, и еще один интервьюируемый подтверждает это: «Наше рассуждение в большей степени становится основанным на управлении, в то время как наши бюджеты становятся недостаточными, а затраты увеличиваются. Все время мы должны отчитываться за деньги, которые мы потратили, и это заставляет продвигать то, за что мы заплатили».

Опрос оценил последствия этого противоречия несколькими способами. В одном примере доля соединений библиотек для всех посещений контента открытого доступа на платформе Revues.org оказалась минимальной. Анкетирование в реальном времени читательской аудитории Revues.org подтверждает эти результаты. Кроме того, оценка наличия и качества установки указательных столбов к услугам Revues.org в документации к системе информационной технологии библиотек показывает значительное неравенство между учреждениями. Обратное справедливо для платных ресурсов, которые систематически оцениваются и продвигаются.

Отсутствие координации между библиотеками и ресурсами открытого доступа, обнаруженное в опросе Эммы Бестер, показывает три сферы трудностей:

- Это плохо для читателей, которые получают руководство и помощь для платных ресурсов, но оставлены без помощи при использовании их собственных устройств – например, поисковой системы Google, – что касается ресурсов открытого доступа. Следовательно, библиотеки перестают действовать в качестве посредников, и их роль разрушается.
- Это плохо для производителей ресурсов открытого доступа, которые не получают выгоды от той же самой поддержки, финансовой поддержки, естественно, а также в терминах обозримости и развития ресурсов, которые библиотеки обеспечивают для других закупок ограниченного доступа.
- Это также плохо для библиотек, которые рискуют стать маргинальными в рамках новой появляющейся документальной экосистемы открытого доступа. Если бюджеты перераспределяются и практики изменяются с использованием других форм посредничества, тогда будущее университетских библиотек находится под угрозой.
- Ситуация, следовательно, становится более ясной. Два препятствия к доминирующей модели распространения исследования открытого доступа, а именно, хрупкость экономических моделей и их взаимосвязь с первичными задачами университетских библиотек, являются частью одной и той же проблемы. Вопрос финансов, вопрос использования и вопрос характера и роли участников экосистемы тесно связаны. Во Франции Центр открытого электронного издательства недавно предложил новую экономическую модель публикаций в сфере гуманитарных и общественных наук, чтобы рассмотреть эту трехстороннюю проблему. Эта модель называется OpenEdition Freemium.

Модель OpenEdition Freemium

С момента основания платформы Revues.org в 1999 г. [4] группа Cleo Центра открытого электронного издательства http://cleo.cnrs.fr/ разрабатывает платформы и услуги электронного издательства открытого доступа для гуманитарных и общественных наук.

Группа Cleo финансируется четырьмя французскими научными организациями и учреждениями высшего образования: CNRS (Национальный центр научных исследований), Школа высшего образования по общественным наукам (EHESS), Университет Прованса и Университет Авиньона. Она также получает финансирование от двух французских киберинфраструктур: TGE Adonis и TGIR BSN (цифровая научная библиотека). Она находится в Марселе и имеет офисы в Париже и Лиссабоне (Португалия). Cleo предлагает целый ряд услуг по электронному издательству, давая возможность

научному сообществу, преподавателям и студентам в области гуманитарных и общественных наук, а также людям вообще, получить доступ к согласованному корпусу электронных ресурсов, опубликованных, главным образом, в открытом доступе. Платформа OpenEdition, http://www.openedition.org, начавшая действовать в начале 2011 г., предлагает один единственный пункт доступа ко всем этим ресурсам, которые распространяются через три платформы: Revues.org с более 330 журналами, всего свыше 80 тыс. документов открытого доступа; Calenda, распространяющая более 18 тыс. научных программ конференций по всем дисциплинам гуманитарных и общественных наук; и Hypothesis.org, приводящая к более чем 350 научным блогам.

После 10 лет развития служб электронного издательства с участием университетов, научных групп и университетских издателей [5] группа Cleo стремилась расширить это сотрудничество до включения библиотек. Цель была в том, чтобы дать возможность библиотекам продвинуть обозримость ресурсов открытого доступа и вовлечь их в более активную роль по финансированию публикаций открытого доступа.

Вот почему в 2011 г. группа Cleo изобрела новую программу, названную OpenEdition freemium. Программа основана на бесплатной экономической модели, используемой многими Интернет-компаниями в настоящее время. Программа Freemium впервые была объявлена идущим на риск капиталистом Фредом Уилсоном на его блоге в 2006 г. и сильно популяризирована журналистом Крисом Андерсеном в его книге "Free: The future of a radical price" [6]. "Freemium" как составное слово, обозначающее «бесплатный» ("free") доступ к некоторым услугам с разрешением других «поощрений» ("premium"). Использование модели Freemium в платформе OpenEdition сочетает бесплатный доступ к информации (например, полнотекстовым статьям и книгам в открытом доступе) с разрешением поощрительных услуг для библиотек [7].

В основном это происходит из-за различия между форматами: полный текст доступен через HTML, тогда как другие форматы, которым обычно отдают предпочтение читатели профессиональной литературы, такие как PDF и ePub, ограничены для подписчиков – пользователей библиотек. Дополнительные услуги для библиотек и их пользователей пакетируются со следующим предложением:

• *Поощрительный доступ к журналам и книжным сериям*. Библиотеки, подписывающиеся на OpenEdition Freemium, предлагают своим пользователям возможность скачивать файлы статей и книг в формате PDF и ePub без ограничений и без цифрового управления правами (DRM) журналов и книжных серий, которые приняли модель Open Access Fremium. Такое скачивание возможно с сайта каждого журнала, а также с использованием платформы Revues.org Bookserver, позволяющей простое скачиивание файлов в формате ePub на мобильные средства, такие как читающие устройства, планшеты и смартфоны.

• *Совокупность прикладных процессов*. Кроме перечня охвата, отслеживания состояния фондов и обеспечения статистики посещений сервера, подписка на OpenEdition Freemium обеспечивает университетскую статистику в соответствии со стандартом COUNTER norm. Эти статистические данные обновляются ежедневно.

• *Экспорт записей в форматах MARC*. Каталог журналов и книг, доступный на платформе Revues.org, был создан более дружественным пользователю путем преобразования его описаний в форму, доступную в форматах UNIMARC и MARC21. Библиотеки, подписывающиеся на OpenEdition Freemium, могут автоматически включать эти записи в свои интегрированные библиотечные управленческие системы (ILMS) путем скачивания файлов ISO2709 или использования сервера Z39.50.

• *Сетевая услуга Calenda*. Пользователи имеют доступ к базе данных Calenda, используя простые запросы, относящиеся ко всем доступным областям (даты, места, категории, участники события и т.д.). Запросы используют сетевую услугу Calenda, - прикладной программный интерфейс (API), который создает потоки данных в нескольких форматах (ATOM, JSON, RSS, iCal). Они могут быть выведены на дисплей и могут быть найдены на сетевых сайтах всех типов, подобных собственным цифровым рабочим пространствам учреждений, на документальных платформах и сетевых страницах факультетов или лабораторий. Таким образом, можно представить на экране все события, происходящие в определенном университете, городе или месте. Также можно фильтровать события по дисциплинам или типу и т.д. Библиотека имеет право сделать открытым использование этой услуги для всего университета или учреждения, которому она принадлежит, становясь посредником этого продвинутого средства для Calenda.

• *Сигнальное оповещение и подписка*. OpenEdition также предлагает услугу сигнального оповещения, позволяющую пользователям автоматически получать сообщения, основанные на выбранных ими ключевых словах, по электронной почте. Услуга уведомляет пользователя каждый раз, когда выбранное выражение появляется на платформах Revues.org, Calenda или Hypotheses.org. Услуга пользуется высоким спросом, и есть возможность устанавливать фильтры, запрашивая информацию, например, только от платформы Calenda. Также можно уточнить поиск, ограничив его определенными полями, запросом сигнального оповещения для статей, опубликованных определенным автором, или ограничить поиск только до названия и реферата. Сигнальное оповещение ограничено до 3 документов на пользователя. Подписчики на OpenEdition Freemium имеют неограниченный доступ.

• *Помощь*. Библиотеки и университеты, которые подписываются на OpenEdition Freemium, имеют прямой доступ по электронной почте и телефону к техническому помощнику, который отвечает на любые вопросы, относящиеся к функционированию платформы и состоянию фондов.

• *Обучение*. Библиотеки, подписывающиеся на OpenEdition Freemium, таже получают пользу от электронного обучения библиотечного персонала на месте на французском или английском языках, включая представление ресурсов платформы, программных средств и новых характеристик. Группа Cleo регулярно организует обучение электронному издательству, использованию программного средства Lodel и научному блогированию в целом. Персонал подписывающихся библиотек имеет преимущественный доступ к летнему университету Cleo по открытому электронному издательству, принимающему каждые два года в Марселе около 100 участников. Тема этого года – «Распространение знания в цифровой век: альянс между авторами, издателями, библиотекарями и читателями относительно цифровой книги». <http://leo.hypotheses.org/5851>.

• *Управление.* Библиотеки, выбирающие бесплатные услуги, могут участвовать в деятельности пользовательского комитета Cleo. Комитет состоит из издателей журналов и книжных серий на платформе Revues.org, издателей научных блогов на платформе Hypotheses.org, сотрудничающих учреждений и библиотек платформы Calenda. Пользовательский комитет собирается раз в год и постоянно занят работой по развитию через рабочие группы, действующие на месте или на расстоянии в таких областях, как цитирование и библиометрия, интернационализация, библиографические средства, качество метаданных и, среди прочего, взаимосвязь с библиотеками. Члены пользовательского комитета получают отчет Cleo о ежегодной деятельности (на французском языке).

Две трети дохода от подписок библиотек предназначается тем журналам и сотрудничающим издателям, которые принимают модель Freemium. Еще одна треть дает возможность Cleo развивать платформу. Весь доход, созданный OpenEdition Freemium, затем снова инвестируется в развитие научного издательства открытого доступа. К 2013 г. 85 научных журналов и 20 издателей, обеспечивающих 1200 книг, будут участвовать в программе OpenEdition Freemium. Благодаря гранту "Equipex" от французского правительства Cleo сможет до 2020 г. перевести в цифровую форму и распространить через OpenEdition Freemium 15 тыс. книг по каждой дисциплине гуманитарных и общественных наук и на нескольких языках. Грант поможет Cleo разработать новые услуги, а также интернационализировать свою платформу. В 2012 г. 32 библиотеки по всему миру подписались на OpenEdition Freemium. Ценовая модель для библиотек является весьма классической. Она основана на ВВП на душу населения страны и количестве студентов или персонала учреждения, подписывающегося на данное предложение. Предложение и его ценовая модель обговариваются с библиотечными консорциумами, такими как Couperin во Франции или Crepuq в Канаде.

ЗАКЛЮЧЕНИЕ

Электронное издательство отчаянно ищет новые экономические модели [8]. «Информация хочет быть бесплатной» - это, кажется, должно стать лозунгом века Интернета. Однако известное высказывание Стюарда Бренда должно быть услышано полностью: «С одной стороны, информация хочет быть дорогóй, поскольку она очень ценная. Именно правильная информация в правильном месте изменяет вашу жизнь. С другой стороны, информация хочет быть бесплатной, поскольку стоимость ее получения все время становится ниже и ниже. Таким образом, вы имеете эти два, сражающиеся друг с другом принципа». Конечно, напряжение между этими двумя принципами оказывает влияние на бизнес-модели в издательской индустрии. Но оно также влияет на роль информационных посредников, таких как библиотеки, которые оценивают информацию. OpenEdition Freemium является предложением рассмотреть противоречие Бренда в ограниченном секторе издательства без отказа от таких традиционных участников, как издатели и библиотеки. В таком случае платформа OpenEdition не пытается вытеснить их, а скорее помогает им создать альянс для бесплатного распространения знания в новой цифровой среде.

ЛИТЕРАТУРА

1. *Willinsky J.* The access principle: The case for open access to research and scholarship. 1er éd. – The MIT Press, Cambridge, MA, 2009. – P. 211.

2. *Morris S.* The true costs of scholarly journal publishing// Learned Publishing. - 2005. – Vol.18, No. 2. – P.115-126.

3. *Bester E.*, *Mounier P.* Usages des resources en libre accès dans les bibliothèques universitaires et services communs de documentation. Le cas Revues.org / Ch. Boukacem-Zegmouri (ed.)//L'information scientifique dans l'univers numérique. Mesures et usages. – ADBS editions, Paris, 2010. – P. 211-226.

4. Dacos M. Revues.org, online humanities and social sciences portal // ELPUB2008. Open scholarship: Authority, community, and sustainability in the age of Web2.0 // Proceedings of the 21th International Conference on Electronic Publishing held in Toronto, Canada 25-27 June 2008. – Leslie Chan and Susanna Mornati, 2008. – P. 426-427. – http://elpub.scix.net/data/works/att/426_elpub2008.content.pdf.

5. *Cavallo D.* Revues.org: L'invention de l'édition électronique scientifique, entre libre accès et modèle économique pérenne // Mémoires du livre. – 2009. – 1, no.1. – http://id.erudit.org/iderudit/038638ar.

6. *Anderson C.* Free: The future of a radical price. – Hyperion, New york City, NY, 2009.

7. *Peyssard J-C.* OpenEdition Freemium: Developing a sustainable library-centered economic model for open access // IFLA Conference, August 13, 2011, Puerto Rico. – http://conference.ifls.org/past/ifla77/164-peyssard-en.pdf.

8. *Dacos M.*, *Mounier P.* L'édition électronique. – Editions La Découverte, 2010. – P. 27.

Москва, ВИНИТИ РАН
28–30 ноября 2012 года

***8-я Международная конференция* «НТИ-2012»,**
посвященная 60-летию ВИНИТИ РАН

«АКТУАЛЬНЫЕ ПРОБЛЕМЫ ИНФОРМАЦИОННОГО ОБЕСПЕЧЕНИЯ НАУКИ, АНАЛИТИЧЕСКОЙ И ИННОВАЦИОННОЙ ДЕЯТЕЛЬНОСТИ»

Для участия **в «НТИ-2012»** приглашаются специалисты в области информационных технологий, телекоммуникаций, создатели и потребители информационных продуктов и услуг, ученые и специалисты РАН, вузовской и отраслевой науки, работники информационных центров и библиотек, служб распространения информационных продуктов и услуг.

Доклады или тезисы докладов, которые будут опубликованы в специальном сборнике, направлять в Оргкомитет конференции «НТИ-2012».

Планируется проведение пленарных заседаний, круглых столов и работа по секциям.

Адрес: Россия, 125190, Москва, ул. Усиевича, 20, Всероссийский институт научной и технической информации (ВИНИТИ РАН), ОНИИР, Оргкомитет **«НТИ-2012».**
Тел: (495) 155-44-22, 155-44-29, 152-64-41,
Факс: (495) 152-54-92, 943-00-60
E-mail: conf@viniti.ru , market@viniti.ru
http://www.viniti.ru

Приглашаем российских и зарубежных авторов к сотрудничеству
в журнале «Международный форум по информации».
Оригинальные статьи и другие материалы (рецензии, письма)
можно присылать на русском или английском языке
по почтовому адресу, указанному в «Памятке для авторов»,
или по электронной почте: mfi@viniti.ru.

Ответственный за выпуск *Л. В. Кобзева*

Компьютерная верстка *М. А. Филимонова*

ИД № 04689 от 28.04.2001 г.
Подписано в печать 06.09.2012 г.
Бумага “Xerox”. Формат 60x841/8. Гарн. литер. Печать цифровая
Усл. печ. л 4,50 Уч.-изд. л. 4,25 Тираж 55 экз.

Адрес редакции: 125190, Россия, г. Москва, ул. Усиевича, д. 20
Тел. (499) 155-44-95