МЕЖДУНАРОДНЫЙ ФОРУМ ПО ИНФОРМАЦИИ

2015 **Том 40** **№ 1**

Выходит ежеквартально **Основан в 1975 г.**

СОДЕРЖАНИЕ

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

К вопросу о семантометрии: новый критерий для оценки вклада научной публикации на основе семантического сходства*

Петр КНОТ
(Petr KNOTH),

Драгомира ГЕРМАННОВА
(Drahomira HERRMANNOVA)

Институт интеллектуальных медиа,
Открытый университет, Великобритания

Предлагается семантометрия (Semantometrics), новый класс метрик для оценки научного исследования. В противоположность существующим библиометрии, вебометрии, альтметрии и т.д. семантометрия не основана на измерении числа взаимодействий в сети научной коммуникации, а посторена на предположении, что для оценки значимости публикации необходим полный текст. Представлено первое семантометрическое измерение, оценивающее научный вклад. Измеряется семантическое сходство публикаций, связанных в сети ссылок, и используется простая формула для оценки их вклада. Мы проводим пробное исследование, в котором проверяем наш подход на небольшом массиве данных и обсуждаем проблемы, возникшие в ходе анализа в существующих массивах данных ссылок. Результаты предполагают, что меры семантического сходства могут быть полезны для обеспечения значимой информации о вкладе научных статей, который не охватывается традиционными мерами влияния, основанными только на ссылках.

ВВЕДЕНИЕ

С тех пор как была внесена идея использования ссылок для научной оценки [1], анализ цитирования получил пристальное внимание, и появилось множество теорий и измерений на основе ссылок. Анализ цитирования стал де-факто стандартом в оценке исследования. Среди основных достоинств оценки статей, основанной на числе ссылок, ими полученных, является простота такого рода измерений и сравнительно хорошая доступность данных цитирования для таких целей.

Тем не менее, ссылки являются одним из многих атрибутов, окружающих публикацию, и сами по себе не обеспечивают достаточное свидетельство влияния, качества и научного вклада. Это происходит из-за широкого ряда характеристик, которые они проявляют, включая изменения отношения (позититвное, негативное), семантику ссылки (сравнение, фактическая информация, определение и т.д.), контекст ссылки (гипотеза, анализ, результат и т.д.) и мотивы цитирования [2], популярность тем и величину научных сообществ [3-4], временную задержку появления ссылок [5], искаженность их распределения [6], различие типов научных статей (теоретическая статья, экспериментальная статья, позиционная, обзор) [4] и, наконец, возможность игры/манипуляции ссылками [7-8].

Последнее десятилетие показало постоянный рост исследований, нацеленных на поиск новых подходов к оценке научных публикаций, таких как вебометрия [9] и альтметрия [10], полагающихся на использование данных вместо ссылок. Традиционное научное направление фокусируется на уменьшении проблем, связанных с использованием ссылок для оценки влияния. Например, недавнее исследование проанализировало распределение ссылок в научных документах относительно традиционной структуры IMRaD (введение, методы, результаты и обсуждение) [11]. Другие исследователи использовали полный текст, чтобы предсказать будущее влияние статей [12] или оценить научные предложения [13]. Авторы работы [14] сравнивали полные тексты и рефераты статей для кластеризации задач и разработали комбинированный подход для отображения научных статей, используя как полный текст, так и классические библиометрические показатели. Параллельно с этой работой движение открытого доступа недавно принесло изменение, дающее возможность иметь свободный доступ и анализировать полные тексты научных статей в широком масштабе, создавая новые направления для развития метрик влияния.

В этой статье представлен подход к оценке влияния статьи, основанный на ее полном тексте (раздел «Гипо-

* Перевод Knoth P., Herrmannova D. Towards Semantometrics: A new semantic similarity based measure for assessing a research publication' s contribution.— http://www.dlib.org/dlib/november14/knoth/11knoth.html

теза»). В этом контексте используется термин *влияние* для обозначения научного вклада в дисциплину, который, по нашему убеждению, является независимым от числа взаимодействий в сети научной коммуникации, но изначально зависит от содержания самой рукописи. Предлагается назвать класс методов, использующих полный текст для оценки научной значимости, *семантометрией*. На основе этого предположения разрабатывалась формула для оценки научного вклада статьи и эмпирически проверялась на небольшом массиве данных (раздел «Эксперимент»). Кроме того, анализируются существующие массивы данных публикаций, описываются их ограничения относительно разработки новых метрик на основе полного текста (раздел «Анализ массивов данных публикаций») и обсуждаются проблемы, возникшие при разработке массивов данных без таких ограничений (раздел «Эксперимент»).

ГИПОТЕЗА

Наша гипотеза основана на предположении, что полный текст публикации необходим для оценки ее влияния. Хотя это может звучать тривиально, но, по нашим сведениям, на данный момент ни одна автоматизированная метрика не использует полный текст. Наша гипотеза утверждает, что добавочная стоимость публикации p может оцениваться на основе семантического расстояния между публикациями, цитируемыми p, и публикациями, цитирующими p. Эта гипотеза основана на понимании того, как исследование полагается на существующее знание, чтобы создать новое знание, на котором другие могут что-то построить. Публикация, которая таким образом создает «мост» между тем, что мы уже знаем, и чем-то новым, которое люди будут развивать на основе этого знания, вносит вклад в науку. Публикация имеет высокий вклад, если она создает «длинный мост» между более дистанцированными областями науки.

С учетом этих идей разработана формула, оценивающая вклад публикации, основывающаяся на измерении семантического расстояния между публикациями, цитируемыми p, и публикациями, цитирующими p.

$$Вклад(p) = \frac{\bar{B}}{\bar{A}} \cdot \frac{1}{|B| \cdot |A|} \sum_{a \in A, b \in B, a \neq b} dist(a,b)$$

Числитель и знаменатель первой дроби формулы подсчитывается по следующему уравнению:

$$\bar{X} = \begin{cases} 1, & \text{если } |A|=1 \text{ или } |B|=1, \\ \dfrac{1}{|x|(|x|-1)} \sum\limits_{\substack{x_1, x_2 \in X \\ x_1 \neq x_2}} dist(x_1, x_2), & \text{если } |A|>1 \text{ и } |B|>1. \end{cases}$$

В формуле B является множеством публикаций, цитирующих публикацию p, и A является множеством, цитируемым p (рис. 1). Сумма в уравнении используется для подсчета общего расстояния между всеми сочетаниями публикаций в множествах A и B. Ожидается, что это расстояние оценивается с помощью использования измерений семантического сходства на полных текстах публикаций, такое как косинусное сходство векторов документов tf-idf. На данный момент мы экспериментируем с рядом этих методов. Вторая дробь в уравнении является фактором нормализации, приспособленным ко всем комбинациям между членами множеств A и B, что приводит к среднему расстоянию между членами этих двух множеств. Первая дробь в выше приведенном уравнении является еще одним фактором нормализации, отвечающим за приспособление значения вклада к отдельной области и типу публикации. Он основан на измерении среднего внутреннего расстояния публикаций в рамках множеств A и B.

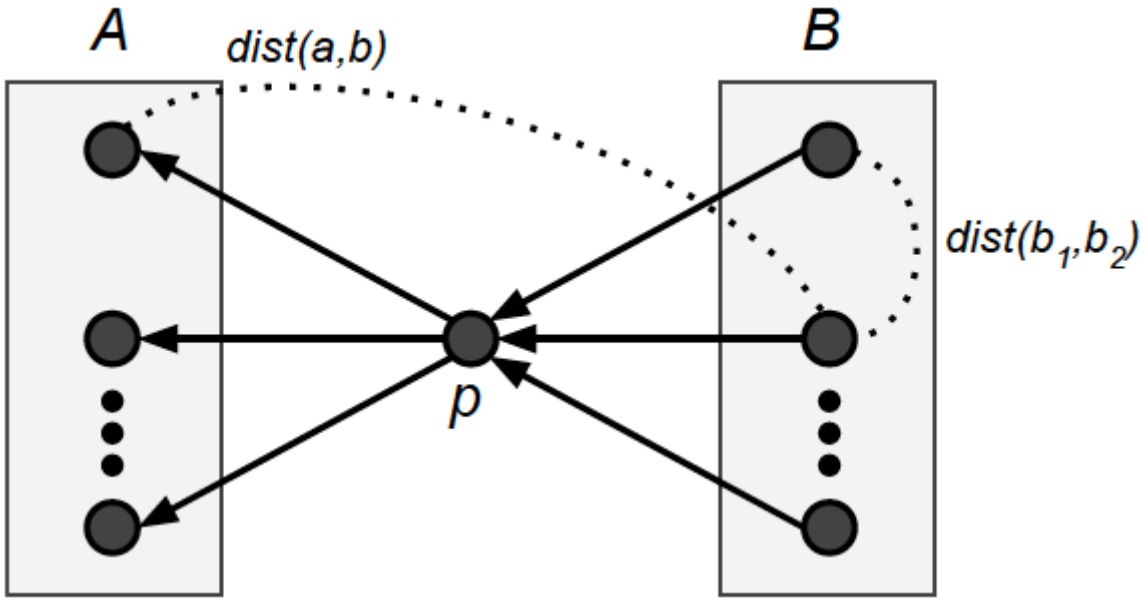

Рис. 1. Объяснение подсчета влияния статьи (p)

Основная идея заключается в том, что, например, в случае обзорной статьи вполне естественно, что публикации внутри множества *A*, а также внутри множества *B* будут распределяться достаточно далеко друг от друга. Однако это не служит признаком вклада статьи, а скорее является естественной особенностью обзорной статьи. С другой стороны, считается, что если статья заимствует идеи из узкой области, но имеет влияние на весьма большую область, то это служит признаком вклада статьи. В обоих случаях первая дробь формулы соответственно признает значение данной метрики.

На практике наш метод оценки вклада статьи означает, что статья с высоким влиянием не нуждается в широком цитировании, однако, ей необходимо инспирировать изменения в своей области или даже определять новую область. Это может проявляться в изменениях словаря, которые являются результатом определенной публикации. Следовательно, слишком активные научные споры относительно обзорной статьи в определенной предметной области, порождающей множество ссылок, будут иметь влияние ниже, чем статья, разрабатывающая новое направление исследования. Важная особенность этой идеи состоит в том, что наш метод не требует столь длительной задержки в оценке, как широко применяемые подсчеты ссылок (как правило, десятки лет) и поэтому может применяться также к сравнительно молодым ученым. Им (методом) трудно манипулировать, он с уважением относится к тому, что научные сообщества различаются по размерам в отдельных дисциплинах, не фокусируется на количестве публикаций, как h-индекс, а скорее – на качественных аспектах. Эксперименты с такими измерениями в прошлом не были возможны, так как по нашим сведениям не было массива, сочетающего информацию по ссылкам с доступом к полным текстам статей.

АНАЛИЗ МАССИВОВ ДАННЫХ ПУБЛИКАЦИЙ

Чтобы проверить нашу гипотезу, было необходимо получить массив данных, удовлетворяющий следующим критериям:

Доступность полного текста служит предпосылкой для проверки нашей гипотезы, так как подсчет сходства требует наличия этой информации.

Плотность сети ссылок относится к доле ссылок и связи ссылок, для которой можно найти статьи и доступ к полным текстам. Эти требования оказалось трудно удовлетворить. Чтобы провести репрезентативную проверку нашей гипотезы, было необходимо убедиться, что наш массив данных содержит значительную долю статей, цитирующих публикацию, а также документы, цитируемые этой публикацией. Если среднее число ссылок на статью приближается к 40 [15], тогда полное множество публикаций, необходимое для оценки вклада одной публикации, должно состоять из 80 публикаций (можно ожидать, что среднее число полученных ссылок будет примерно таким же, как ее среднее число ссылок). Если бы нам хотелось изучить вклад 100 публикаций, нам потребовалось бы множество из ≈8000 статей. Получение такого множества представляется затратным по времени из-за ограничений машинного доступа к публикациям и прав доступа по подписке.

Многодисциплинарность является важной в силу предположения, что передаваемое знание между различными научными областями служит показанием научного вклада публикации. А значит массив данных для проверки нашей гипотезы должен содержать значительную долю статей, цитируемых этой публикацией в условиях оценки, а также статьи, ссылающиеся на эту публикацию и в первую очередь статьи из различных предметных областей.

Наше первоначальное ожидание состояло в том, чтобы найти подмножество публикаций, удовлетворяющее всем нашим критериям в рамках области открытого доступа. По этой причине использовался массив данных CORE [16], который обеспечивает доступ к научным статьям, собранным из архивов и журналов открытого доступа. Однако, поскольку многие цитирования и ссылки все еще относятся к подписке (никакой контент открытого доступа CORE не мог быть законно собран), мы решили, что сеть ссылок является слишком разбросанной для целей нашей оценки. Тем не менее, верится, что ситуация вскоре улучшится благодаря ратифицированным во многих странах мира мандатам правительств, требующим издавать финансируемые государством исследования через открытый доступ. Следовательно, эксперимент проводился на укрупненном массиве данных, автоматически загружающем отсутствующие в открытом доступе документы с сетевых сайтов издателей. Однако выяснилось, что эта задача представляется трудновыполнимой из-за широкого ряда ограничений, наложенных издателями на машинный доступ к публикациям (даже открытого доступа), находящимся в их системах [17].

Поэтому мы проанализировали несколько других массивов данных, в частности Open Citation Corpus, ACM dataset и DBLP+Citation dataset. Хотя эти массивы предоставляют данные из различных научных дисциплин, ни один из них не содержит полные тексты публикаций и по этой причине не может быть использован.

Более того, изучались массивы KDDcup dataset и iSearch collection. Оба массива являются подмножествами раздела по физике архива ArXiv database и содержат полные тексты статей и ссылки. Массив KDD dataset включает около 29 тыс. документов, а iSearch collection – 150K документов в формате PDF. К сожалению, оба массива охватывают только одну дисциплину, что делает их непригодными для нашего эксперимента.

ЭКСПЕРИМЕНТ

Поскольку ни один существующий массив данных не подходил для нашей задачи, мы решили создать новое небольшое множество, удовлетворяющее всем вышеуказанным критериям. Этот массив данных создан вручную с отбором 10 ядерных публикаций массива CORE с уровнем варьирования ссылок в системе Google Scholar. Статьи, цитируемые этими публикациями, и ссылки этих публикаций, которые отсутствовали в CORE, были загружены вручную и добавлены к нашему массиву данных. Включались только документы, для которых мы нашли свободно доступную онлайн версию. Публикации, написанные не на английском языке, удалялись из массива данных, так как наш метод подсчета сходства не был разработан для множества языков. Приведенная ниже таблица содержит список 10 публикаций с числом загруженных документов, написанных на английском языке. Всего нам удалось загрузить 62% всех документов, найденных в качестве непосредственных соседей ядерных документов в сети цитирования. После удаления не англоязычных статей массив был сокращен до 51% от всей сети цитирования. Весь процесс занял два дня, и окончательный массив документов в итоге (http://core.kmi.open.ac.uk/contribution_dataset) содержал 716 документов в формате PDF.

Таблица

Массив данных и результаты эксперимента

№ п/п	Название статьи	Авторы	Год издания	$\|B\|$ (оценка ссылки)	$\|A\|$ (число ссылок)	Вклад
1	Open access and altmetrics: distinct but complementary	Ross Mounce	2013	5 (9)	6 (8)	0,4160
2	Innovation as a Nonlinear Process, the Scientometric Perspective, and the Specification of an "Innovation Opportunities Explorer"	Loet Leydesdor, Daniele Rotolo and Wouter de Nooy	2012	7 (11)	52 (93)	0,3576
3	Ranking of library and information science researchers: Comparison of data sources for correlating citation data, and expert judgments	J.A. Li, et al.	2010	12 (20)	15 (31)	0,4874
4	The Triple Helix of university-industry-government relations	Loet Leydesdor	2012	14 (27)	27 (72)	0,4026
5	Search engine user behaviour: How can users be guided to quality content?	Dirk Lewandowski	2008	16 (30)	12 (21)	0,5117
6	Revisiting h measured on UK LIS and IR academics	M. Sanderson	2008	25 (41)	8 (13)	0,4123
7	How journal rankings can suppress interdisciplinary research: A comparison between Innovation Studies and Business & Management	Ismael Rafols, et al.	2012	39 (71)	70 (128)	0,4309
8	Web impact factors and search engine coverage	Mike Thelwall	2000	53 (131)	3 (10)	0,5197
9	Web Science: An Interdisciplinary Approach to Understanding the Web	James Hendler, et al.	2008	131 (258)	22 (32)	0,5058
10	The Access/Impact Problem and the Green and Gold Roads to Open Access: An Update	Steven Harnad, et al.	2004	172 (360)	17 (20)	0,5004
				474 (958)	232 (428)	

Примечание: Документы ранжированы по их оценкам ссылок. Цифры за скобками представляют число англоязычных документов, которые были успешно загружены и обработаны. Цифры в скобках отражают размер всего массива. Последняя колонка показывает оценку вклада. Массив данных может быть загружен с сайта http://core.ac.uk/contribution_dataset

Мы обработали эти статьи используя программное обеспечение CORE и вычислили оценку вклада для ядерных документов. Это было проделано в два этапа:

1. Извлекался текст из всех документов PDF с помощью библиотеки извлечения текста (Apache Tika).
2. Вычислялась оценка вклада с помощью использования косинусной меры сходства на векторах tf-idf [18], созданных из полных текстов в качестве средств подсчета оценки вклада.

Точнее, расстояние, используемое в оценке вклада, было подсчитано как $dist(d_1,d_2)$=1- $sim(d_1,d_2)$, где $sim(d_1,d_2)$ – косинусное сходство документов d_1 и d_2, а значение (1- $sim(d_1,d_2)$ часто рассматривается как расстояние, хотя это не метрика чистого расстояния, так как она не удовлетворяет свойству треугольного неравенства).

Результаты для каждого из 10 документов можно посмотреть в таблице. Интересно отметить, что наблюдаются весьма значительные различия между оценкой вклада публикаций и очень похожих оценок ссылок. Более пристальный анализ показал, что наш подход помогает эффективно отфильтровывать самоцитирования в похожей работе или, что точнее, вызывает меньше доверия к ним. Также публикация с самой высокой оценкой ссылки не имеет самой высокой оценки вклада, фактически ее оценка вклада ниже, чем у публикации, цитируемой в 10 раз меньше. Мы утверждаем, что это указывает на то, что публикации с достаточно низкой оценкой ссылки все еще могут обеспечить высокий вклад в науку.

На рис. 2 показано сравнение оценки вклада с оценкой ссылки и числом ссылок. Линия на каждом из рисунков показывает линейную модель соответствия. Можно наблюдать, что оценка вклада медленно растет с увеличением числа ссылок. Это ожидаемое поведение, так как вероятность, что публикация обеспечивает высокий вклад в число тем и дисциплин, усиливается с подсчетом ссылок, однако они не прямо пропорциональны. Например, в массиве данных можно найти публикации с оценкой ссылок ниже и относительно высокой оценкой вклада. Это демонстрирует, что даже публикации с низкой оценкой ссылок могут обеспечивать высокий вклад в науку. С другой стороны, с увеличением числа ссылок оценка вклада медленно снижается. Это отражает, что рост числа цитируемых документов не может непосредственно воздействовать на оценку вклада.

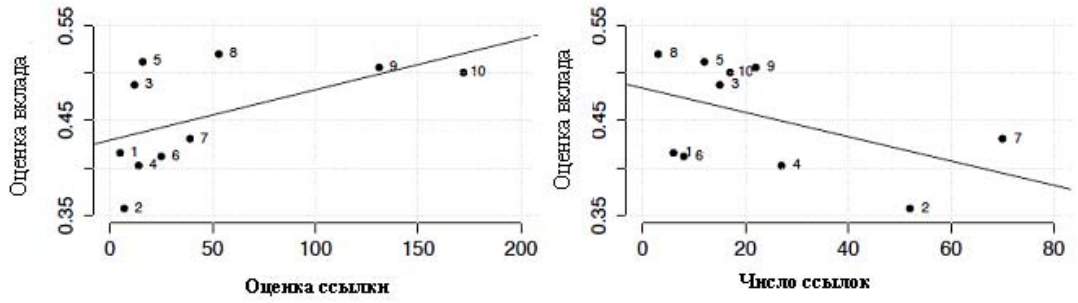

Рис.2. Сравнение оценок вклада с оценками ссылок и с числом ссылок

ОБСУЖДЕНИЕ И ВЫВОДЫ

Использование современных метрик эффективности научных публикаций (библиометрии, альтметрии, вебометрии и т.д.), по нашему мнению, основано на ложном предположении, что влияние (или даже качество) научной статьи может оцениваться исключительно на основе внешних данных без учета самой рукописи публикации. Такое предположение напоминает идею оценки судебного разбирательства без сомнения относительно выступления в суде подозреваемого, а значит таким же образом являет слабым. Мы показали, что новые измерения влияния, принимающие в расчет рукопись публикации, могут быть разработаны. Считаем, что эта идея предлагает много возможностей для исследования этого класса измерений, называемого нами семантометрией. Результаты нашего пробного исследования показывают, что мера, основанная на семантическом сходстве публикаций в сети цитирования, является перспективной и должна анализироваться дальше на большем массиве данных.

Более того, мы продемонстрировали важность развивающихся массивов данных, на которых этот класс мер может быть проверен, и объяснили проблемы их разработки. Первоначальный вопрос – проблема распределения данных ссылок, которая является естественным следствием работы по цитированию публикаций из различных дисциплин и баз данных. Так как системы, создаваемые организациями, имеющими обусловленные соглашения с издателями, такими как Google Scholar, не обмениваются данными, существует потребность в провайдерах открытого доступа для объединения сил, создающих один массив данных, распространяющийся на все научные дисциплины. В общем мы верим, что эта ситуация демонстрирует потребность в поддержке открытого доступа к научным публикациям не только для того, чтобы люди читали, но также для доступа машин. Эта проблема особенно современна, так как исключение в законе об авторском праве в Великобритании по извлечению текста было недавно одобрено парламентом этой страны и вступило в силу в июне 2014 г. [19, 20], создавая новые возможности для развития инновационных услуг.

ЛИТЕРАТУРА

1. *Garfield E.* Citation indexes for science. A new dimension in documentation through association of ideas// Science. — 1955. —Vol. 122, No.3159 (October).— P.108-111. — http://dx.doi.org/10.1126/science.122.3159.108

2. *Nicolaisen J.* Citation Analysis// Annual Review of Information Science and Technology. — 2007.— Vol. 41, No.1.— P. 609-641. —http://doi.org/10.1002/aris.2007.1440410120

3. *Brumback R. A.* Impact factor wars: Episode V — The Empire Strikes Back// Journal of Child Neurology. — 2009. — Vol. 24, No. 3 (March). — P.260-262.— http://doi.org/10.1177/0883073808331366

4. *Seglen P. O.* Why the impact factor of journals should not be used for evaluating research// BMJ: British Medical Journal. — 1997.— No. 314 (February).— P. 498-502. — http://dx.doi.org/10.1136/bmj.314.7079.497

5. *Priem J., Hemminger B. M.* Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web//First Monday. —2010. — Vol.15, No.7 (July).

6. *Seglen P. O.* The skewness of science// Journal of the American Society for Information Science. — 1992.— Vol. 43, No.9 (October).— P. 628-638.

7. *Arnold D. N., Fowler K. K.* Nefarious numbers// Notices of the American Mathematical Society. — 2010.— Vol. 58, No.3.— P. 434-437.

8. *The PLoS Medicine Editors.* The impact factor game// PLoS medicine.— 2006.—Vol. 3, No.6 (June).— http://dx.doi.org/10.1371/journal.pmed.0030291

9. *Almind T. C., Ingwersen P.* Informetric analyses on the world wide web: Methodological approaches to "webometrics"// Journal of Documentation. —1997. — Vol. 53, No. 4. — P.404-426.— http://dx.doi.org/10.1108/ EUM0000000007205

10. *Priem J., Taraborelli D., Groth P., Neylon C.* Altmetrics: A manifesto. — 2010.

11. *Bertin M., Atanassova I., Lariviere V., Gingras Y.* The distribution of references in scientific papers: An analysis of the IMRaD structure// Proceedings of the 14th ISSI Conference.— 2013.— Vienna, Austria— P. 591-603.

12. *Yan R., Huang C., Tang J., Zhang Y., Li X.* To better stand on the shoulder of giants//Proceedings of the 12th Joint Conference on Digital Libraries, pages 51-60, Washington, DC. — ACM. 2012.— http://dx.doi.org/10.1145/ 2232817.2232831

13. *Holste D., Roche I., Hörlesberger M., Besagni D., Scherngell T., Francois C., Cuxac P., Schiebel E.* A concept for inferring "Frontier Research" in research project proposals// Proceedings of the 13th ISSI.—2011. —Durban, South Africa— P. 315-326.

14. *Glenisson P., Glanzel W., Persson O.* Combining full-text analysis and bibliometric indicators. A pilot study// Scientometrics. —2005.— Vol. 63, No.1.— P.163-180. — http://doi.org/10.1007/s11192-005-0208-0

15. *Abt H. A., Garfield E.* Is the relationship between numbers of references and paper lengths the same for all sciences? //Journal of the American Society for Information Science and Technology.— 2002.— Vol. 53, No.13.— P.1106-1112. —http://dx.doi.org/10.1002/asi.10151

16. *Knoth P., Zdrahal Z.* Core: Three access levels to underpin open access// D-Lib Magazine— 2012.—Vol. 18, No. 11/12. — http://doi.org/10.1045/november2012-knoth

17. *Knoth P., Rusbridge A., Russell R.* Open mirror feasibility study, Appendix A: Technical prototyping report//Jisc report.— 2014.

18. *Manning C.D., Raghavan P., Schütze H.* An introduction to information retrieval. — Cambridge University Press, online edition.— 2009.

19. *Hargreaves I.* Digital opportunity: A review of intellectual property and growth.— Technical report.— 2011.

20. *Intellectual Property Office.* Implementing the Hargreaves review.— 2014.

На плечах гигантов: растущее влияние старых статей*

Алекс ВЕРСТАК
(Alex VERSTAK),

Анураг АЧАРИА
(Anurag ACHARYA),

Хелдер СУЗУКИ
(Helder SUZUKI),

Син ХЕНДЕРСОН
(Sean HENDERSON),

Михаил ЯХЬЯЕВ
(Mikhail IAKHIAEV),

Клиф Чиунг Ю ЛИНЬ
(Cliff Chiung Yu LIN),

Намит ШЕТТИ
(Namit SHETTY)

Google Inc.

В статье изучается эволюция влияния более старых научных статей. Делается попытка ответить на четыре вопроса. Первый – как часто старые статьи цитируются в научных статьях и как это меняется со временем. Второй – как влияние старых статей варьируется в различных научных областях. Третий - ускоряется или замедляется изменение влияния старых статей. Четвертый - отличаются ли эти тенденции в отношении очень старых статей. Чтобы ответить на эти вопросы, изучались ссылки из статей, опубликованных в 1990-2013 гг. С помощью компьютера для исследования была вычислена доля ссылок на старые статьи из ежегодно публикуемых статей в качестве критерия влияния. Для этого исследования статьи, опубликованные, по меньшей мере, за 10 лет до появления цитирующей статьи, рассматривались как старые статьи. Чтобы изучить, как изменения в поведении цитирования различаются по областям исследования, были вычислены эти цифры для 261 предметной категории и 9 широких областей исследования. В конце проведено повторное вычисление для двух других определений старых статей – 15 и 20 лет и более. Имеются три основных вывода, полученные в результате нашего исследования. Первый – влияние старых статей в значительной степени выросло за период 1990-2013 гг. Анализ показывает, что в 2013 г. 36% ссылок было на статьи, возраст которых не меньше 10 лет, и эта доля возросла на 28 % с 1990 г. Доля старых ссылок за 1990-2013 гг. увеличилась для 7 из 9 широких областей исследования и для 231 из 261 предметной категории. Второй – изменение за вторую половину (2002-2013 гг.) было значительно большим, чем за первую (1990-2001 гг.), рост во второй половине удвоился по сравнению с ростом в первой половине. Третий – тенденция роста влияния старых статей также относится и к статьям, которым не менее 15 и 20 лет. В 2013 г. 21 % ссылок был на статьи возраста ≥ 15 лет с ростом в 30 % с 1990 г., а 13 % ссылок относились к статьям возраста ≥ 20 лет с ростом в 36 % за тот же период. Теперь, когда нахождение и чтение релевантных старых статей является таким же легким делом, как и нахождение и чтение недавно опубликованных статей, важные достижения не теряются на полках библиотек и повсеместно влияют на работу спустя годы.

ВВЕДЕНИЕ

В течение двух последних десятилетий наблюдался ряд серьезных изменений в научной коммуникации. Во-первых, научные журналы в большей степени перешли от физического распространения печатных номеров к доступности отдельных статей в режиме онлайн. Большое количество журналов также провело оцифровку старых статей и сделало их доступными в режиме онлайн. Во-вторых, поисковые службы теперь индексируют полный текст статей, а не только рефераты и ключевые слова. Подход общего расположения продвинулся от прямо противоположного хронологического порядка (в первую очередь самые последние статьи) к расположению по релевантности (в первую очередь наиболее релевантные статьи). В-третьих, теперь многие

* Перевод Verstak A., Acharya A., Suzuki H., Henderson C., Iakhiaev M., Lin C.Ch. Yu., Shetty N. On the shoulders of giants: The growing impact of older articles. – http://arxiv.org/pdf/1411.0275v1.pdf

журналы сделали статьи более быстро доступными, часто вскоре после их принятия. Кроме того, ряд дисциплин создал большие массивы препринтов, включающие статьи до их принятия для формальной публикации. Следовательно, ученые могут узнать о новых результатах гораздо раньше, чем это было возможно прежде. В-четвертых, количество статей и журналов быстро растет, за период 1990-2013 гг. число ежегодно публикуемых научных статей выросло почти в три раза. В результате имеется гораздо больше последних работ, чтобы ученые могли что-то узнавать из них, основываться на них и цитировать их.

Первые два изменения облегчают ученым процесс нахождения наиболее релевантных статей для их работы, независимо от возраста статей. Нахождение и чтение релевантных старых статей является в настоящее время таким же легким, как и нахождение и чтение недавно опубликованных статей. Если бы это были единственные изменения, то было бы резонным ожидать, что доля ссылок на старые статьи увеличится.

Второй набор изменений значительно расширил количество конкурентных и недавно вышедших работ, так что ученым необходимо разместить свою работу соответственно этому. Если бы это были единственные изменения, то было бы обоснованным ожидать, что доля ссылок на последние статьи увеличится, а доля ссылок на старые статьи снизится.

Чтобы понять эволюцию влияния старых научных статей, мы изучили ссылки из статей, опубликованных в 1990-2013 гг., и попытались ответить на четыре вопроса. Первый – как часто старые статьи цитируются в научных статьях и как это со временем меняется. Второй – как влияние старых статей варьируется в различных областях науки. Третий – ускоряется или замедляется изменение влияния старых статей. Четвертый - являются ли эти тенденции разными для более старых статей.

С помощью компьютера мы вычислили для исследования долю ссылок на старые статьи из ежегодно публикуемых статей в качестве критерия влияния. Для данного исследования мы определили статьи, опубликованные не менее 10 лет назад до появления цитирующей статьи, как *старые статьи*. Чтобы изучить, как изменения в поведении цитирования различаются в областях исследования, мы вычислили эти цифры для 261 предметной категории и 9 широких областей исследования. В конце мы повторили вычисление для двух других определений старых статей – 15, 20 лет и более.

Три основных вывода вытекают из нашего исследования. Первый – влияние старых статей, как измерено ссылками, значительно выросло в 1990-2013 гг. Анализ показывает, что в 2013 г. 36% ссылок было на статьи, возраст которых 10 лет и больше, и что эта доля выросла на 28% с 1990 г. Несмотря на то, что в эволюции имелись отклонения, связанные с областью исследования, доля старых ссылок увеличилась в 1990-2013 гг. для 7 из 9 широких областей исследования и для 231 предметной категории из 261.

Второй вывод - изменение за вторую половину (2002-2013 гг.) было значительно большим, чем за первую (1990-2001 гг.), рост во второй половине удвоился по сравнению с первой. В нашем контексте большинство усилий по оцифровке архивов, а также сдвиг в сторону поиска на основе полного текста, ранжированного по релевантности, произошли во второй половине (2002-2013 гг.).

Третий вывод – тенденция к росту влияния старых статей также касается статей, имеющих возраст 15 лет и более, и статей с возрастом от 20 лет и более. В 2013 г. 21% ссылок относился к статьям ≥ 15 лет с ростом в 30% с 1990 г., а 13% ссылок было на статьи ≥ 20 лет с ростом в 36% за тот же самый период.

Сейчас, когда нахождение и чтение релевантных старых статей является таким же легким, как нахождение и чтение недавно опубликованных статей, важные достижения не теряются на полках и повсеместно влияют на работу, несмотря на прошедшие годы.

В оставшейся части данной статьи, в целях краткости, мы называем ссылки на старые статьи *старыми ссылками*. В следующем разделе мы описываем этапы анализа. Далее представляем и обсуждаем результаты. После этого мы описываем родственные, связанные с нашим исследованием работы.

МЕТОДЫ

Для проведения исследования мы включили все журналы и конференции, приписанные к одной или более категориям в документе Scholar Metrics 2014 года. Критериями включения публикаций в Scholar Metrics были [1,2]: 1)публиковать 100 или более статей в 2009-2013 гг. 2)по крайней мере одна статья должна получить хотя бы одну ссылку в 2009-2013 гг., 3)следовать рекомендациям Google Scholar в отношении индексирования. Scholar Metrics ограничивает категоризацию до предметных категорий для публикаций на английском языке. Соответственно данное исследование охватывает все журналы на английском языке и конференции, включенные в Scholar Metrics. Scholar Metrics отображает до 20 публикаций высшего ранга на одну предметную категорию и делает оставшиеся публикации доступными через поиск по ключевым словам. Данное исследование охватывает *все* категоризированные журналы и конференции, а не только 20 публикаций верхнего ранга на предметную категорию. Scholar Metrics также включает хранилища отобранных препринтов. Хранилища препринтов в данном исследовании не отражены.

Мы использовали все предметные категории, общим числом 261, из документа Scholar Metrics 2014 г. Чтобы изучить направления в широких областях, мы также сгруппировали предметные категории в 9 широких научных областей. Для этого использовали широкие области из Scholar Metrics, сделав лишь одно изменение – мы разделили *Engineering and Computer Science* (Машиностроение и Вычислительная наука). Модели цитирования в этих двух областях очень различаются, и такое разделение позволило нам изучить эти различия. Кроме того, мы добавили пункт *All articles* (Все статьи) для объединения всех широких областей.

Мы создали группу статей для каждой комбинации предмет–категория-год и широкая–область–год, такую как *Immunology* (Иммунология) за 2000 г. или *Physics and Mathematics* (Физика и Математика) за 2004 г. Каждая группа категория–год/область–год включала все статьи, опубликованные в определенный год во всех публикациях в определенной категории/области.

Для каждой публикации мы включили все статьи с датой публикации в рамках 1990-2013 гг. Заметим, что каждый журнал или конференция могут ассоциироваться с более, чем одной предметной категорией. Такие публикации включены в вычисления для каждой категории, частью которой они являются.

Для каждой группы категория–год/область-год мы вычислили общее число ссылок, а также число ссылок на статьи, опубликованные в каждый предшествующий год. Такие подсчеты ссылок (общее число ссылок, а также число ссылок за каждый предшествующий год) включали *все* ссылки из этих статей, а не только ссылки

на статьи, включенные в данное исследование. Мы использовали эту матрицу для вычисления доли ссылок на старые статьи. Использовались три различных порога для старых статей - ≥ 10 лет, ≥ 15 лет и ≥ 20 лет.

Чтобы увидеть, набирает ли темп или снижается скорость изменения в доле старых ссылок, мы вычислили общее изменение для периода 1990-2001 гг. (первая половина) и для периода 2002-2013 гг. (вторая половина) по каждой категории.

РЕЗУЛЬТАТЫ

На рис. 1 представлена эволюция доли ссылок на статьи, которым не менее 10 лет. На нем отражены все публикации, включенные в исследование. Показано, что доля старых ссылок устойчиво растет в 1990-2013 гг. Он также демонстрирует, что темп роста резко остановился в 1990-1999 гг., а после этого ускорился.

На рис. 2. отображена эволюция доли старых ссылок для всех широких областей. Он показывает, что 7 из 9 широких областей имели существенный рост в доле старых ссылок за период 1990-2913 гг. Две широкие области – *Chemical and Material Sciences* (Химия и Материаловедение) и *Engineering* (Машиностроение) не претерпели значительного изменения в доле старых ссылок.

В табл. 1 представлена доля старых ссылок, а также изменение с 1990 г. в цифровой форме (для простоты сравнения). Изменение за 1990-2013 гг. вычислено в виде процентного отношения:

(*доля_в*_2013 - *доля_в*-1990)/*доля_в*_1990 * 100

Таблица показывает, что в 2013 г. четыре широкие области имели по меньшей мере 40% ссылок на старые статьи, *Гуманитарные науки и Искусство* были на уровне 51%. Пять широких областей имели рост свыше 30% в доле старых ссылок в период 1990-2013 гг., самый большой рост – 56% произошел в *Бизнесе, Экономике и Управлении.*

Табл. 2 представляет собой гистограмму различий в доле старых ссылок для каждой из широких категорий. Колонки в гистограмме подсчитывают число предметных категорий, рост которых находится в рамках установленного диапазона. Таблица показывает, что в общей сложности 231 категория из 261 (89%) наблюдала рост в доле старых ссылок. То есть рост в доле старых статей произошел в разных областях, которые очень различаются в терминах свой частоты публикации и моделей цитирования.

Если посмотреть более пристально, она показывает, что 102 предметные категории из 261 наблюдали рост в доле старых ссылок, который составлял свыше 50%. Доля двух широких областей – *Бизнес, Экономика и Управление* и *Вычислительная наука,* чуть меньше, чем две трети предметных категорий (10 из 16 и 11 из 18, соответственно) наблюдали рост более 50% в доле старых ссылок. Она также показывает, что все широкие области имели несколько предметных категорий с > 50 % роста в доле старых ссылок. Наконец, она показывает, что большая доля предметных категорий, претерпевших падение в доле старых ссылок, является частью *Химии и Материаловедения* и *Машиностроения.*

Заметим, что некоторые предметные категории включены в более, чем одну категорию. Например, *Экономика развития, Людские ресурсы* и *Организации* являются частью как *Бизнеса, Экономики* и *Управления*, так и *Общественных наук.* В результате, суммарный подсчет в колонке в табл. 2, то есть число предметных категорий, которые наблюдали рост в 0-20%, будет большим, нежели число в той же колонке для *всех статей.*

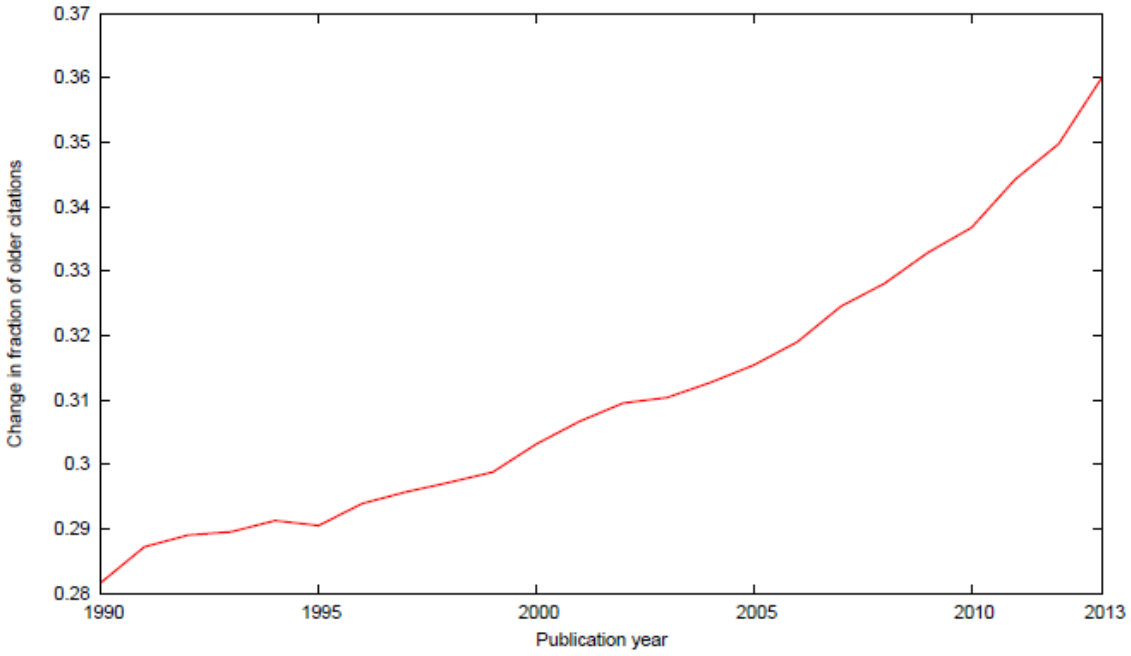

Рис. 1. Доля старых ссылок из всех статей, опубликованных в 1990-2013 гг.

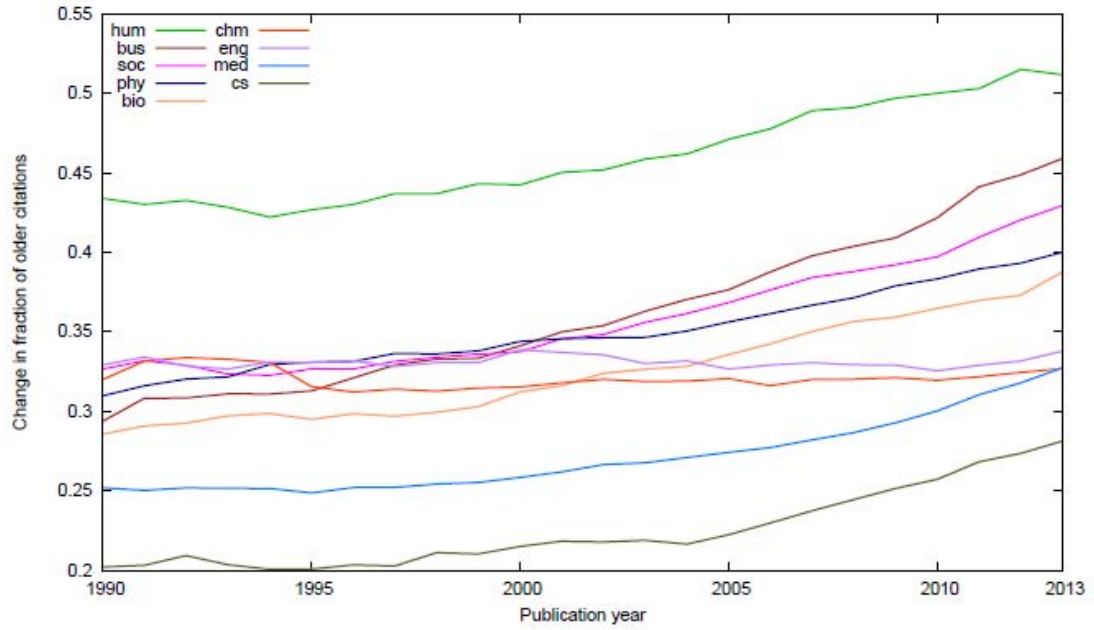

Рис. 2. Доля ссылок на старые статьи для широких областей исследования

Примечание: **bio -** Life Sciences and Earth Sciences (Науки о жизни и о Земле); **bus –** Business, Economics and Management (Бизнес, Экономика и Управление); **cs –** Computer Science (Вычислительная наука); **chm** – Chemical and Material Sciences (Химия и Материаловедение); **eng** – Engineering (Машиностроение); **hum** – Humanities, Literature and Arts (Гуманитарные науки, Литература и Искусство); **med** – Health and Medical Sciences (Здравоохранение и Медицина); **phy** – Physics and Mathematics (Физика и Математика); **soc** – Social Sciences (Общественные науки).

Таблица 1

Изменение в доле старых ссылок в 1990-2013 гг.

Широкая область	Старые ссылки в 2013 г.	Изменения с 1990 г.
Гуманитарные науки, литература и искусство	51%	18%
Бизнес, экономика и управление	46%	56%
Общественные науки	43%	31%
Физика и математика	40%	29%
Науки о жизни и о Земле	39%	36%
Машиностроение	34%	3%
Химия и материаловедение	33%	2%
Здравоохранение и медицина	33%	30%
Вычислительная наука	28%	39%
Все статьи	36%	28%

Таблица 2

Гистограмма изменения в доле старых ссылок для широких областей

Широкая область	<0%	0-20%	20-30%	30-40%	40-50%	>50%
Гуманитарные науки, литература и искусство	2	10	5	5	2	2
Бизнес, экономика и управление	0	0	3	1	2	10
Общественные науки	4	10	11	8	9	10
Физика и математика	2	3	10	4	2	3
Науки о жизни и о Земле	6	10	9	7	2	5
Машиностроение	10	15	4	7	2	2
Химия и материаловедение	11	5	1	0	0	2
Здравоохранение и медицина	3	30	17	8	2	9
Вычислительная наука	1	2	2	0	2	11
Все статьи	30	73	56	36	22	44

Примечание: Подсчет в каждой колонке является числом предметных категорий, рост которых находится в рамках установленного.

Изменение в темпе роста

Табл. 3 показывает изменение в доле старых ссылок за 1990-2001 гг. и 2002-2013 гг. Заметим, что основной линией для отсчета роста всех процентных отношений в этой таблице является доля старых ссылок в 1990 г. Использование общей основной линии отсчета позволяет сравнить рост процентных отношений непосредственно.

Таблица показывает, что повсеместно рост во второй половине был немного больше, чем двойное увеличение в первой половине. Для всех широких областей, имевших нетривиальный рост в доле старых ссылок в 1990-2013 гг., увеличение во второй половине было значительно большим, чем рост в первой половине. Для 6 из 9 областей рост во второй половине увеличился по меньшей мере вдвое по сравнению с первой половиной.

Что можно сказать о даже более старых статьях?

На рис. 3 показано изменение в доле старых ссылок для двух других определений слова «старая» - по меньшей мере 15 и 20 лет. На рисунке видно, что доля ссылок даже на более старые статьи росла постоянно в 1990-2013 гг. и что рост во второй половине (2002-2013 гг.) был значительно выше, чем рост в первой половине (1990-2001 гг.).

На рис. 4 отображена эволюция доли ссылок на более старые статьи для всех широких областей. Показано, что 7 из 9 широких областей наблюдали значительный рост в доле ссылок даже на более старые статьи в период 1990-2013 гг. Две широкие области, *Химия* и *Материаловедение,* а также *Машиностроение*, не претерпели значительного изменения.

Таблица 3

Изменение в доле ссылок на старые статьи в 1990-2001 гг. и 2001-2013 гг.

Широкая область	Изменение в 1990-2001 гг.	Изменения в 2002-2013 гг.
Гуманитарные науки, литература и искусство	4%	14%
Бизнес, экономика и управление	19%	37%
Общественные науки	5%	26%
Физика и математика	11%	18%
Науки о жизни и о Земле	11%	25%
Машиностроение	2%	1%
Химия и материаловедение	-1%	3%
Здравоохранение и медицина	4%	26%
Вычислительная наука	8%	31%
Все статьи	9%	19%

Примечание: Основной линией отсчета для всего роста процентных отношений является доля старых ссылок в 1990 г.

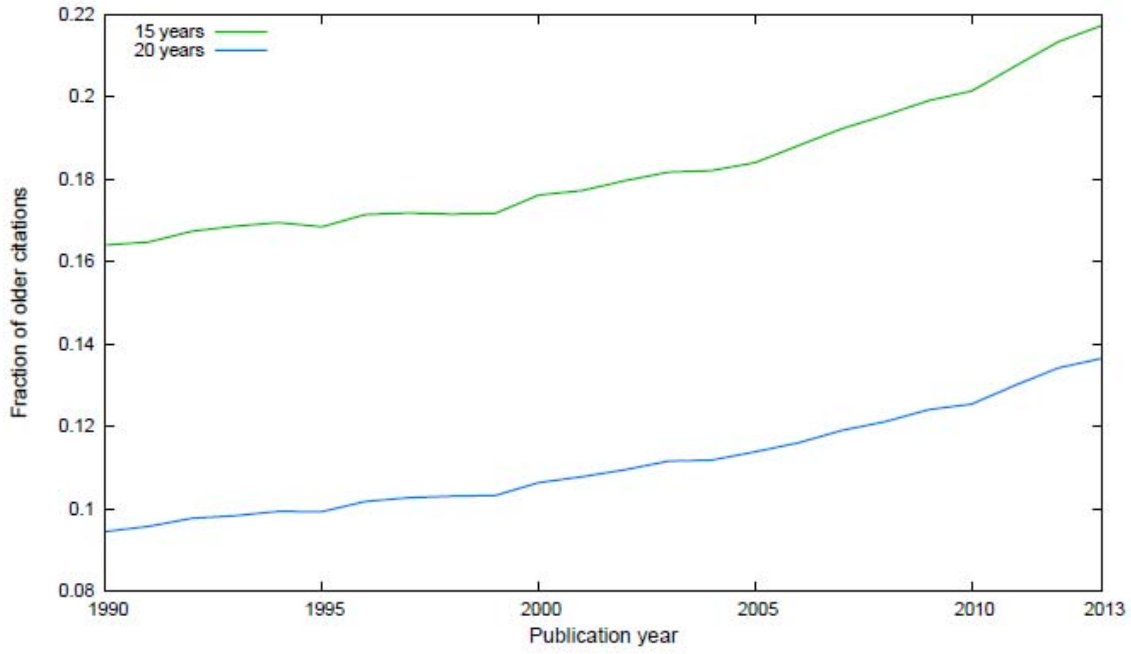

Рис. 3. Доля ссылок даже на более старые статьи во всех публикациях

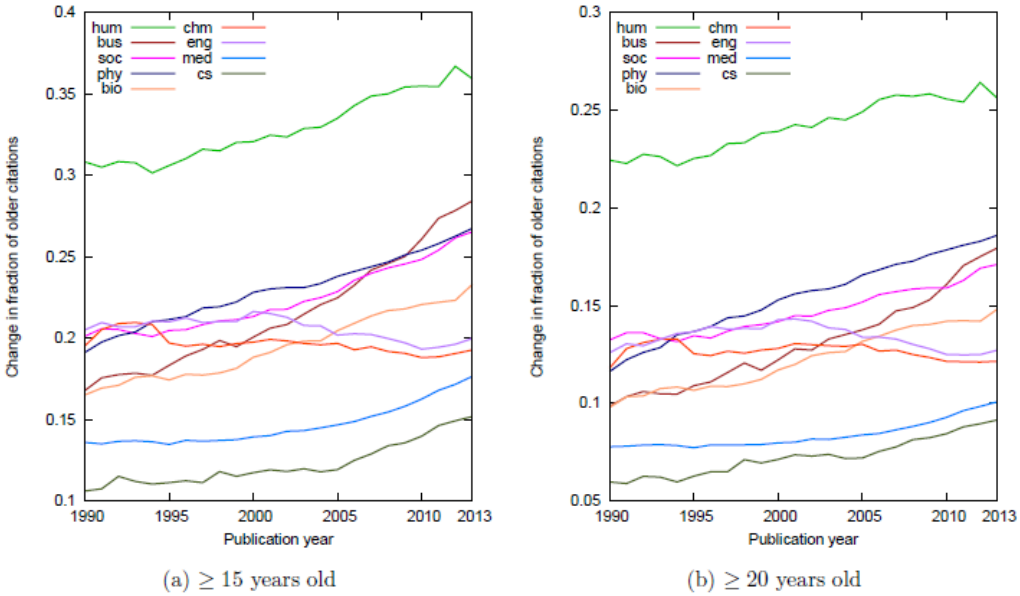

Рис. 4. Доля ссылок даже на более старые статьи для широких областей исследования

Примечание: Расшифровку сокращений hum, chm, bus, eng, soc, med, phy, cs, bio см. в рис. 2.

СВЯЗАННАЯ С ИССЛЕДОВАНИЕМ РАБОТА

Исследование возраста ссылок в научных статьях имеет давнюю историю. Первоначальная работа в данной области изучала возраст ссылок в качестве способа измерения темпа «устаревания» (obsolescence) в научной литературе [3-8]. Одной из задач изучения устаревания было обеспечить руководство для библиотек, касающееся политики сохранения старых томов журналов. Ранние метрики для возраста ссылок включали «полураспад» (half-life) [3] и "Price's Index" (Указатель Прайса) [9]. Лайн [5] указал на потенциальный эффект роста в ряде статей, опубликованных по вопросу метрик, касающихся возраста ссылок – если число статей растет быстро, то можно ожидать, что доля ссылок на недавно опубликованные статьи будет расти так же быстро.

Изучая понятие устаревания скорее с точки зрения использования, а не перспективы цитирования, Сендисон [10] обнаружил, что после начального периода использование старых выпусков журналов по физике в Массачусетсском технологическом институте не снизилось в связи с их возрастом.

В ранней статье, изучающей потенциальное влияние доступа в режиме онлайн на научную коммуникацию, Одлижко [11] сообщил, что после начального периода частота доступа к статьям в режиме онлайн из ряда массивов не изменялась с возрастом статей. Основываясь на этом, он говорил, что легкий доступ в режиме онлайн к оцифрованным фондам, становящимися доступными, должен привести к более широкому использованию старых материалов.

Не так давно Иванс [12] исследовал влияние доступности в режиме онлайн журнальных статей относительно возраста ссылок. На основе анализа указателей ссылок издательства Thomson Reuters и базы данных Information Today Inc. относительно доступности журнальных статей в режиме онлайн он сделал вывод, что поскольку больше журнальных выпусков выходят в режиме онлайн, то статьи, на которые ссылаются, имеют тенденцию быть более свежими. Он говорил, что сдвиг от просмотра печатных массивов к поиску в массивах онлайн облегчает аннулирование старой литературы.

Эти результаты напрямую контрастируют с нашими. Мы не нашли свидетельства того, что доступность в режиме онлайн ведет к сокращению ссылок на старые статьи. Наоборот, мы выявили, что для большинства областей рост в доле старых ссылок ускоряется в тот период, когда существенное число статей становится доступным в режиме онлайн и их можно находить по полному тексту.

Этим результатам также противоречат два других исследования, опубликованные примерно в то же время, что и работа Иванса [12], и которые применяли другие подходы к анализу. Хантингтон и др. [13] изучали модели использования статьи на основе протоколов сетевого доступа для журнальных массивов OhioLink. Они выявили, что имелось два этапа в истории доступности научных статей. Первый этап охватывал от 8 до 9 лет с даты публикации. В этот период использование часто снижалось, наиболее острым снижение было в первые 2-3 года (на треть за первый год и примерно на 60% к третьему году). Следующий этап обычно имел относительно устойчивый уровень использования. Анализируя HTTP Referer headers на предмет запросов относительно журнальных статей, они обнаружили, что пользователи поисковых служб больше хотели смотреть старые статьи, чем пользователи просмотровой среды. Они говорили, что это различие происходит из-за подхода, применяющего ранжирование релевантности, который используется службами поиска в сети.

Ларивье и др. [14] изучали ссылки из большого массива статей, опубликованных в 1900-2004 гг. Они пришли к выводу, что полезная жизнь научных публикаций постоянно увеличивается с 1970-х гг. И что в совокупности всех научных дисциплин и в области естественных наук и машиностроения в частности, доля старых ссылок постоянно растет. Наши результаты согласуются с результатами Ларивье и др.

ЗАКЛЮЧЕНИЕ

Три основных вывода вытекают из нашего исследования. Первое – влияние старых статей, как измерено с помощью ссылок, значительно росло в период 1990-2013 гг. Наш анализ показывает, что в 2013 г. 36% ссылок приходились на статьи, возраст которых 10 лет, и что эта доля выросла на 28% с 1990 г. Доля старых ссылок увеличилась в 1990-2013 гг. для 7 из 9 широких областей исследования и для 231 предметной категории из 261.

Второе – для большинства областей изменение во второй половине (2002-2013 гг.) было значительно большим, чем в первой (1990-2001 гг.). Везде рост во второй половине был в два раза больше роста в первой половине. Заметим, что большая часть усилий по оцифровке архивов, а также сдвиг в сторону поиска на основе полного текста и ранжирования релевантности произошли во второй половине.

Третье - тенденция растущего влияния старых статей также касается статей, возраст которых 15 лет и 20 лет. В 2013 г. 21% ссылок был на статьи с возрастом ≥ 15 лет с ростом в 30 % с 1990 г., а 13% ссылок относились к статьям с возрастом ≥ 20 лет с увеличением в 36%.

Во «Введении» мы упомянули два широких направления, обладающие потенциалом влиять на долю старых ссылок. Во-первых, нахождение и чтение релевантных старых статей сейчас происходит так же легко, как и нахождение и чтение недавно опубликованных статей. Это облегчает ученым процесс цитирования наиболее релевантных статей для их работы, независимо от возраста статей. Во-вторых, наблюдался резкий рост числа статей, опубликованных за год. Это значительно увеличило число последних статей и ученым надо разместить свою работу с учетом ее цитирования.

Наши результаты предполагают, что из этих двух направлений легкость нахождения и чтения наиболее релевантных статей, независимо от их возраста, имеет большее влияние. Для большинства областей ретроспективная оцифровка, как и включение в широко распространенную поисковую службу с ранжированием по релевантности, произошла во второй половина периода исследования. Как упоминалось ранее, это также и период, когда наблюдался большой рост в доле старых ссылок.

В настоящее время, когда нахождение и чтение релевантных старых статей является таким же простым, как нахождение и чтение недавно опубликованных статей, важные достижения не становятся потерянными на полках и повсеместно влияют на работу даже по прошествии многих лет.

ЛИТЕРАТУРА

1. Google Scholar Mettrics help page. – http://scholar.google.com/intel/en/scholar/metrics.html, 2014.

2. *Suzuki H.* 2014 Scholar Metrics Released. – http://googlescholar.blogspot.com/2014/06/2014-scholar-metrics-released.html, 2014.

3. *Burton R.E., Kebler R.W.* The half-life of some scientific and technical literatures // American Documentation. – 1960. — Vol. 11, No. 1. – P. 18-22.

4. *Lawler E.E.* III. Psychology of the scientist: IX. Age and authorship of citations in selected psychological journals // Psychological Reports. – 1963. – Vol. 13, No. 2. – P. 537-537.

5. *Line M.B.* The "half-life" of periodical literature: Apparent and real obsolescence // Journal of Documentation. – 1970. – Vol.26, No. 1. – P. 46-54.

6. *Line M.B., Sandison A.* Progress in documentation: "Obsolescence" and changes in the use of literature with time //Journal of Documentation. – 1974. Vol. 30, No. 3. – P. 283-350.

7. *Meadows A.J.* The citation characteristics of astronomical research literature // Journal of Documentation. – 1967. - Vol. 23, No. 1. – P. 28-33.

8. *Oliver M.R.* The effect of growth on the obsolescence of semiconducter physics literature // Journal of Documentation. – 1971. – Vol. 27, No. 1. – P. 11-17.

9. *Price D.J.* Citation measures of hard science, technology, and nonscience // Communication among scientists and engineers. – 1970. – P. 3-22.

10. *Sandison A.* Densities of use, and absence of obsolescence, in physics journals at MIT //Journal of the American Society for Information Science. – 1974. – Vol. 25, No. 3. – P.172-182.

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

12. *Evans J.A* Electronic publication and the narrowing of science and scholarship // Science. – 2008. – Vol. 321, No. 5887. – P. 395-399.

13. *Huntington P., Nicholas D., Jamali H.R., Tenopir C.* Article decay in the digital environment: An analysis of usage of OhioLINK by date of publication, employing deep log methods // Journal of the American Society for Information Science and Technology. – 2006. – Vol. 57, No. 13. – P. 1840-1851.

14. *Larivière V., Archambault E., Gingras Y.* Long-term variations in the aging of scientific literature: From exponential growth to steady-state science // Journal of the American Society for Information Science and Technology. – 2008. – Vol. 59, No. 2. – P. 288-296.

Информационные потребности ученых в сфере библиографических баз данных: обзор литературы*

Моргана Карнейро де АНДРАДЕ

(Morgana Carneiro de ANDRADE)

Докторская программа в области технологии и информационных систем, Университет провинции Минью, Португалия

Ана Алисе БАПТИСТА

(Ana Alice BAPTISTA)

Научный центр алгоритмов, Отделение информационных систем, Университет провинции Минью, Португалия

Статья представляет собой обзор литературы, цель которого определить существующие информационные потребности ученых, когда они обращаются к библиографическим базам данных. Первоначально было найдено 192 статьи с использованием баз данных Scopus, Web of Science и Google Scholar. После применения критериев для исключения число статей сократилось до 16, что уже является показателем небольшого числа исследований на эту определенную тему. Результаты показывают, что определить информационные потребности ученых сложно. Они также свидетельствуют, что ученым требуется информация с высокой степенью гранулирования (детализации). Мы приходим к выводу, что хотя доступные исследования обеспечивают важную информацию об информационных потребностях ученых и дают совет, как их рассматривать, существует необходимость в более глубоких исследованиях. Результаты таких более глубоких исследований могут быть полезны, чтобы служить в качестве показания к созданию новых процедур и средств, включая те, которые основаны на элементах новых метаданных, полученных для улучшения поисковых результатов с помощью средств Linked Open Data tools.

ВВЕДЕНИЕ

Тема информационных потребностей появилась с начала проведения исследований в области библиотековедения и документации, а затем информатики. С появлением Интернета наблюдался рост исследований на эту тему, особенно с концентрацией внимания на информационных службах, таких как цифровые библиотеки и библиографические базы данных.

Согласно авторам [1], «понимание информационных потребностей, поведения, связанного с поиском информации, и использования информации учеными требует внимания» и становится более сложным, поскольку ученые играют несколько ролей (исследователь, преподаватель, администратор и т.д.), их потребности и интересы меняются со временем, и они постоянно подвергаются влиянию технологического прогресса.

Огромный объем информации в Интернете и разнообразие услуг стали парадоксально препятствовать идентификации наиболее релевантных статей. Чтобы найти релевантную информацию за короткий срок, от пользователя требуется, чтобы он знал «[…] что получить, где получить и как получить» [2]. Эти вопросы соответствуют тому, что в области библиотековедения и информатики называется информационной потребностью [2].

Одним из предвестников в этой области был Тейлор [3], который в статье «The Process of Asking Question» («Процесс задавания вопроса») пролил свет на понимание информационных потребностей, его мысли остаются уместными по сей день. Автор предлагает четыре уровня информационных потребностей:

- первый уровень – осознанные и неосознанные потребности, которые, будучи идентифицированными, относятся к «совершенному (точному) вопросу»;
- второй уровень – осознанные потребности, которые плохо определены, но которые будут понятными из взаимодействия с другими людьми;
- третий уровень – осознанные потребности, которые хорошо определены, но которые не могут быть должным образом введены в информационную систему;
- четвертый уровень – осознанные потребности, которые хорошо определены и могут быть «переведены» на язык информационной системы способом, с помощью которого их можно обработать.

Исходя из этих подходов, Тейлор приводит некоторые аспекты, которые влияют на отношение человек-машина: а) организация системы, которая включает ха-

* Перевод de Andrade M.C., Baptista A.A. Information needs of researchers in bibliographic databases environment: A literature review. — http://elpub.scix.net/data/works/att/104_elpub2014.content.pdf

рактеристики ввода и вывода; б) типы, сложности и характеристики предмета, относящегося к вопросу, и в) компетентность ученого.

«Внутренняя организация», часть организации системы и ее характеристик ввода, как это понимает автор, соответствует пунктам доступа, которые, по его мнению, связаны со степенью простоты в использовании терминов, глубины анализа и индексирования и уровня специфичности. Эти пункты доступа должны предполагать «многомерное пространство» протяженностью от эмпирических данных до теоретических понятий, полученное с помощью дескриптивных данных, экспериментального свидетельства, исторического материала, анализа результатов, интерпретации дескриптивных категорий информации. По мнению Тейлора, способ эксплуатации информационной службы имеет причастность к тому, как ученый формирует свои вопросы и какое количество релевантных ответов он получает из системы.

В уровнях Тейлора, со второго по четвертый, присутствует роль библиотекаря в качестве интерпретатора и «переводчика» потребностей пользователя для системы. Благодаря технологическому прогрессу и знакомству ученых с технологией, роль библиотекаря становится менее заметной. Поскольку пользователи не так часто прибегают к помощи библиотекаря, информационные службы должны предложить функциональные возможности, отвечающие потребностям ученых, предоставляя пункты доступа, которые позволяют находить релевантные документы. Поэтому для данного обзора литературы мы сфокусировались на одном из моментов, определенных Тейлором, а именно на том, как внутри организованы, классифицированы и заиндексированы службы, а также на их пунктах доступа. В связи с этим мы считаем, что ученый знает, что он хочет, но зависит от «внутренней организации» информационных служб в смысле удовлетворения своих информационных потребностей.

В этом смысле определение пунктов доступа, отвечающих потребностям пользователя, может быть полезным, чтобы служить показанием к созданию новых процедур и средств, включая те, которые основаны на элементах метаданных, взятых для улучшения поисковых результатов из Linked Open Data tools. Тенденция к увеличению глубины детализации на описанном уровне позволяет, чтобы такие инициативы, как W3C «The RDF data cube vocabulary», «Data Catalog Vocabulary», связанные данные могли быть более легко адаптированы [4,5]. Цыганик и др. [4], публикуя руководящие принципы от инициатив W3C до многомерной публикации данных, подтверждают мысль Тейлора [3], высказанную 52 года назад, поскольку Тейлор также подчеркивал необходимость в многомерном пространстве для эмпирических данных и теоретических понятий. Также согласно Исмаилу и Кариму [6], Всемирная сеть не обеспечивает поддержку начинающим ученым, и поэтому одно из решений может заключаться в том, чтобы информационные службы стали семантически взаимодействующими. Одним из предложений разрешить эти проблемы было использование технологий семантической Всемирной сети.

Данная статья имеет следующую структуру: Описание научного исследования – содержит описание стратегий для проведения поиска. Результаты – представляются и обсуждаются результаты, дающие картину того, что пока разработано в целях удовлетворения информационных потребностей ученых. Заключение – делаются окончательные выводы.

ОПИСАНИЕ НАУЧНОГО ИССЛЕДОВАНИЯ

Мы начали с определения исследований, рассматривающих потребности ученых при обращении к библиографическим базам данных и того, что разработано в целях удовлетворения этой потребности. Для поиска использовались базы данных Scopus, Web of Science, Networked Digital Library of Theses and Dissertations (NDLTD), Library and Information Science, and Technology Abstracts (LISTA) и Google Scholar. Использованными ключевыми словами были: информационные потребности и библиографическая база данных; поведение информационного поиска и научная коммуникация; профиль пользователя и научная информация; научные статьи и потребности пользователя; поиск и библиографическая база данных и потребности пользователя; изучение пользователя и библиографическая база данных; информационные потребности и научная коммуникация.

Указанные выше термины использовались в полях ключевых слов/тем в базах данных, а в Google Scholar использовалось поле названия. При поиске в базах данных мы отобрали следующие временные ограничения: никаких; язык: английский, португальский и испанский. Та же самая процедура использовалась для Google Scholar, но в этом случае анализ результатов ограничивался первыми 100 наиболее релевантными статьями.

Статьи, явившиеся результатом данного поиска, были отобраны на основе названий и рефератов (шаг 1). Статьи, на которые ссылались в этих исследованиях, также были отобраны на основе названий и рефератов (шаг 2). Мы приняли такую же процедуру для отбора статей, цитирующих те статьи, которые были найдены в ходе первого шага (шаг 3). Таким образом мы получили массив с большим охватом, но без потери релевантности теме.

Чтобы сохранить логичность предлагаемого исследования, мы исключили статьи, которые считали находящимися вне сферы изучения, такие как применение программного обеспечения, информационное поведение, анализ релевантности, технический анализ информационно-поисковых систем. Кроме того, мы определили несколько исследований, подход которых был более общим и которые анализировали такие аспекты, как вид используемых источников, язык или предмет [7, 8]. В этом случае мы предпочли не включать эти статьи в результаты, чтобы не исказить релевантную информацию, касающуюся цели данного обзора литературы. Включение было ограничено научными статьями. На основе этих критериев число статей, реально связанных с целью настоящего исследования, было сокращено со 192 до 16.

РЕЗУЛЬТАТЫ

Анализ 16 статей дал картину того, что изучалось относительно информационных потребностей ученых при поиске в библиографических базах данных, сведения представлены в таблице.

Четырьмя основными аспектами, выявленными в ходе анализа, являются: компоненты статей, индексирование и метаданные, область и профиль пользователя.

- *Использование компонентов научных статей в качестве способа повышения результатов поиска*

В контексте данного исследования компоненты статей представляют физические или логические структуры документа [9-11]. Таблицы и рисунки являются примерами физических компонентов, а данные, полученные в результате какого-либо эксперимента, – логическими структурами или изложениями фактов [10].

Таблица

Предметы и аспекты, рассматриваемые в статьях

Аспекты / Авторы	Предмет	Компоненты	Область	Индексирование/ метаданные	Профиль пользователя
Amato & Straccia, 1999	IN				x
Bates, Wilde, & Siegfried, 1993	ISB		x		
Bates,1996	ISB		x		
Bishop, 1998	ISB	x	x	x	
Bishop, 1999	IN	x		x	
Borgman, 1986	IN			x	
Courtright, 2007	ISB				x
Crowston & Kwasnik, 2003	IN			x	
Dogan et al., 2009	IN		x		
Hjørland, Nielsen, &Williams, 2001	ISB	x		x	
Ismail & Kareem, 2011	IN			x	
Lee & Downie, 2004	IN			x	
Markey, 2007a	ISB		x		
Markey, 2007b	ISB		x		
Rowlands, 2007	ISB		x		
Sandusky & Tenopir, 2008	ISB	x		x	x

Примечание: IN – информационные потребности, ISB – поведение при информационном поиске

Бишоп [10] провела исследование, которое изучало, как компоненты научной статьи определяются, хранятся и используются пользователями в цифровых библиотеках. В своем исследовании она использовала DeLIver для того, чтобы позволить ученым Иллинойского университета находить компоненты документов. Бишоп [11] сообщила, что ученые ценят использование определенных компонентов в конкретных ситуациях. Ученые также продемонстрировали важность использования этих компонентов для решения вопроса о том, какие статьи, являющиеся результатом поискового процесса, необходимо читать. Исследование Бишоп, как она говорит, стоит в одном ряду с принципами, которые поддерживал Поль Отле [10,11].

Согласно Отле, аналогично тому, как ученые–химики продвигались от анализа молекул к атомам, усилия должны направляться на то, чтобы подумать о способах, позволяющих науке иметь доступ к определенным частям компонента публикации. В своих мечтаниях, в первых десятилетиях 1900-х гг., Отле сказал: «будут найдены методы для быстрого и полного индексирования работ, чтобы позволить немедленно осуществлять поиск, без хлопот и без труда, той реальной ценности, которую каждая публикация вносит в знание» [12]. Рейворд [13] дополняет, что эти высказывания относятся к атомам информации, которые должны претерпеть изменения, чтобы отвечать информационным интересам и потребностям пользователей.

Йёрланд, Нилсен и Уильямс [14] приводят результаты Бишоп, чтобы осветить «дискуссию относительно необходимости замены традиционных линейных структур в документах свободной комбинацией «инфоблоков», являющихся «извлечением отдельных фактов и идей в качестве самостоятельных единиц» [10]. Авторы также упоминают исследования Аль-Хавамде и др., Аль-Хавамде и Уиллета, Лалмаса и Ратвена, чьи исследования поддерживают использование компонентов текстов.

- *Связь с областью*

Область рассматривается рядом авторов как фактор, который служит помехой информационным потребностям ученых и способу, с помощью которого пользователи действуют во время поиска. Например, можно было бы ожидать, что фасетный поиск, сопряженный с булевыми операторами в базах данных, работающих в режиме онлайн, даст лучшие результаты, независимо от области ученого. Однако, согласно исследованию Бейтса [15] в области гуманитарных наук, конечные пользователи сочли трудным осуществлять такие виды поисков, а также находить оптимальные результаты.

Эта идея подкреплена Марки [16,17], который считает, что эксперты в определенной области стремятся к высокой степени точности, ограничивая область поиска. Их стратегии основаны на идентификации «ключей», содержащихся в любом слове или фразе в названии, фамилии автора и ее вариантах, на испытательном или определенном научном центре, сокращающих число искомых пунктов, но дающих высокую степень релевантности.

Область также связана с модификацией информационных потребностей ученых, например через появление новых, более специализированных научных областей [18]. Примером служит исследование Догана и др. [19], которое обнаружило, что большинство поисков в PubMed в предметной области осуществлено по гену, протеину и/или болезни. Это особенно интересно, если принять в расчет, что всего несколько десятилетий назад вместо гена или протеина проводящие поиск уче-

ные использовали термины, связанные с каким-то видом терапии или диагнозом.

- *Процесс индексирования и пункты доступа*

Информационные потребности пользователей часто формулируются неопределенно не только в том, что касается терминов, но также и в том, что касается структуры системы, в которой проводится поиск. Процесс осуществления поиска включает такие проблемы, как синтаксис, семантика, структура и цель поиска; как используются пункты доступа для того, чтобы сократить и расширить результаты; поисковые альтернативы, если неудача в поиске возникает в результате личной ошибки или ошибки системы [20].

Сандацки и Тенопир [9] утверждают, что существует большая трудность для ученых в том, чтобы определить релевантные статьи, поскольку они все еще индексируются таким образом, что не содержат подробную информацию о документе. Делается ссылка на ProQuest CSA, разработавшую прототип системы, которая обеспечивает подробную информацию с помощью индексирования отдельных компонентов статей. Эта модель позволяет реализацию булевых поисков, используя автора, название, статистику, географические и таксономические термины. Поиски могут быть усовершенствованы с помощью применения карт, рисунков, фотографий, типа статьи или прогнозирующих моделей. Эта процедура связана с растущим уровнем детализации, используемой в проектировании баз данных. Бишоп [10] дополняет утверждения Сандацки и Тенопир констатацией того, что индексирование статей высоко стандартизировано, включая идентификацию автора, название, реферат и ключевые слова. Далее Бишоп заявляет, что имеющиеся в статьях отдельные элементы, не изучающиеся в поисковых областях, могут продвинуть поиск наиболее релевантных документов.

Бейтс, Уайлд и Зигфрид [21] представляют анализ использования библиографических баз данных в режиме онлайн. Эти результаты внесли вклад в совершенствование (1) фасетов тезауруса «The Styles and Periods Art and Architecture Thesaurus» с помощью включения ряда терминов, которые прежде не допускались в этом тезаурусе; (2) структуры базы данных при помощи включения авторов–художников на основе вариантов фамилий и терминов, определяющих академические дисциплины в этой базе данных. Согласно указанным авторам, только качественное индексирование способствует получению высокоточных и релевантных результатов поиска.

Как и Бишоп [10,11], Кроустон и Квасник [22] тоже затронули проблему индексирования и подчеркнули ее связь с решающим фактором – контекстом. Кроустон и Квасник идентифицировали трудность соответствовать потребностям пользователей релевантными результатами, связав ее с такими проблемами, как неточная или неполная информация в базах данных, связанная с индексированием.

Исследования, в которых рассматривается процесс индексирования, все еще находятся в начальной стадии относительно того, что касается возможностей расширения уровня анализа и детализации. Это не соответствует ряду инициатив, в которых использование дескриптивных метаданных перестает быть ограниченным до идентификации названия, автора и предмета. Такие актуальные вопросы, как обработка результата, факторы риска и вывод начинают изучаться либо вручную, либо автоматически, особенно с использованием средств Semantic Web [23, 24].

Йёрланд и др. [14] говорят, что различные структуры, существующие в текстах, влияют на поиск. Поисковая стратегия в научных базах данных может варьироваться в соответствии с определенными пунктами доступа, такими как методологические проблемы или полученные сведения, считающимися темами, представляющими наибольший интерес.

Что касается описания и извлечения материала, то существуют метаданные или пункты доступа, которые являются результатом деятельности индексирования документов, т.е. пункты доступа определяют объективные возможности поиска документа пользователями через алгоритмические и автоматические процедуры [14]. В этом смысле исследования показывают, что агенты программного обеспечения могут обрабатывать информацию, содержащуюся в статьях, используя семантическую обработку в качестве новой формы эксплуатации контекста [25].

Ли и Дауни [26] разработали исследование в отношении информации по музыке (music information retrieval - MIR) в сфере музыкальных цифровых библиотек (music digital library – MDL). Одним из вопросов был следующий: «какие типы метаданных или пунктов доступа должны быть предоставлены пользователям?». Что касается этого вопроса, то они определили потребность в новых типах метаданных в качестве пунктов доступа, которые включают информацию о музыке или музыкальных объектах и информацию, контекстуализирующую поиски реального мира пользователя.

Йёрланд и др. [14] подчеркивают, что предметные пункты доступа (Subject Access Points – SAP) являются крайне необходимыми для поиска документов. Таким образом, если эти пункты доступа не предоставлены информационными службами, то пользователи не могут искать необходимые им документы. Некоторые исследования также показывают, что увеличение детализации информации в информационных службах способствует совершенствованию поиска, поскольку обеспечивает более широкий круг пунктов доступа [9, 10, 14, 15, 20,26]. Ограничивая эти высказывания, мы говорим, что ожидается более высокая степень детализации пунктов доступа, чтобы помочь получить поисковые результаты высокой точности и релевантности, которые в большей степени соответствуют потребностям пользователей. Это связано с заявлениями Отле относительно информационной автоматизации. Как указывает Менцель [27], «выраженные и осознанные желания отдельных лиц в любую эпоху ограничены их восприятием того, что является осуществимым». Относительно того, что в настоящее время является возможным для большинства информационных служб, то сюда не входят поиски на основе высокой детализации и пункты доступа; возвращаясь к 3-му уровню информационных потребностей, изложенному Тейлором, потребности ученых не могут быть должным образом реализованы. Возможно это лучше объясняет, почему все еще недостаточно исследований относительно информационнных потребностей ученых, касающихся библиографических поисков. Отсюда возникает необходимость изучать перспективу ученого, что ему нужно и какие пункты доступа будут отвечать его потребностям.

- *Профиль пользователя*

Проблема отношения к пользователю изложена Кортрайт [28], когда она наблюдала, что имело место изменение в направлении исследований относительно того, что касается информационных потребностей. Обычно исследования концентрировались на модели, в

центре которой находится система, а затем были перенаправлены на модель, в центре которой стоит пользователь, где исследование концентрируется на информации относительно участвующих лиц.

Амато и Страччиа [2] и Кортрайт [28] считают, что информационные потребности могут варьироваться в зависимости от типа пользователя и от контекста, в котором эти потребности анализируются или запрашиваются. Итак, одна из проблем в развитии исследований по информационным потребностям состоит в том, чтобы установить, какой профиль пользователя следует анализировать.

Кроме того, ученые – это пользователи, чей опыт в развитии поисков предполагает информацию с наивысшим уровнем определенности и релевантности. Сандацки и Тенопир [9] обнаружили, что для этого типа пользователя идентификация релевантных документов в короткое время заставляет подумать об определенных типах поведения, таких как, например, время, отведенное на чтение статей. Эти результаты подкреплены исследованиями Бергеля и др. [29], Шоттона [23], Тенопир и др. [30,31].

ЗАКЛЮЧЕНИЕ

Данный обзор литературы представил некоторые интересные и обещающие аспекты, связанные с информационными потребностями ученых и которые сто́ят дальнейшего изучения: компоненты статей, индексирование и метаданные, область и профиль пользователя. Понятно, что Отле действительно был за рамками своего времени в том, что касается его предложений относительно потребности получить доступ к определенным частям документов быстро и в полном объеме. Однако большинство авторов сообщали о трудностях в идентификации информационных потребностей ученых. Возможно, они относятся к трудности, которую создают сами ученые при формулировании своих релевантных потребностей, прибегая к использованию информационных служб, направляющих их к 3-му уровню информационных потребностей, предложенному Тейлором.

Другой аспект, определенный в ходе этого обзора литературы, касается 4-го уровня информационных потребностей, т.е. осознанных, хорошо определенных потребностей и которые могут быть «переведены» на язык информационной системы так, чтобы их можно было обработать. Наблюдается тенденция использовать высокую степень детализации информации, которая может быть оптимизирована с помощью использования семантической Всемирной сети и служб объединенных данных, способных обеспечить более релевантные результаты поиска за меньшее время и с меньшим усилием со стороны ученых.

Что касается развития данного исследования, то стоит отметить обнаруженные нами трудности и ограничения. Некоторые трудности, с которыми мы столкнулись, связаны с рядом проблем, определенных авторами, на которых мы ссылались в данном обзоре литературы. В частности, у нас были проблемы, связанные с процессом индексирования баз данных (включение определенных ключевых слов), который вызвал ограничение в числе искомых статей. Это подразумевало необходимость в адаптации стратегии поиска, чтобы идентифицировать другие релевантные статьи для данного обзора литературы.

В качестве будущей работы мы предлагаем сконцентрировать исследования на информационных потребностях, удовлетворить которые можно было бы с помощью использования технологий семантической Всемирной сети. Например, как ученые могут воспользоваться потенциалом, предложенным этими технологиями? Как эти технологии используются для лучшего соответствия потребностям ученых?

Благодарность. Мы выражаем благодарность федеральному университету Эспирито Санто, Бразилия; Организации CAPES, Министерству образования Бразилии за финансовую поддержку нашей научной деятельности. Часть данной работы финансировалась фондами FEDER через программу COMPETE и национальными фондами через FCT (Foundation for Science and Technology) по проекту FCOMP-01-0124-FEDER-022674.

ЛИТЕРАТУРА

1. *Kuruppu P. U., Gruber A. M.* Understanding the information needs of academic scholars in agricultural and biological sciences// The Journal of Academic Librarianship. —2006. — Vol. 32, No.6.—P. 609–623. — http://www.sciencedirect.com/science/article/pii/S0099133306001510. —doi:10.1016/j.acalib.2006.08.001

2. *Amato G., Straccia U.* User profile modeling and applications to digital libraries//Research and Advanced Technology for Digital Libraries (p. 184–197). — Berlin: Springer, 1999. — http://link.springer.com/chapter/10.1007/3-540-48155-9_13.— doi: 10.1007/3-540-48155-9_13

3. *Taylor R. S.* The process of asking questions// American documentation.—1962. — Vol. 13, No.4. — P. 391–396. — http://onlinelibrary.wiley.com/doi/10.1002/asi.5090130405/abstract.— doi: 10.1002/asi.5090130405

4. *Cyganiak R., Reynolds D., Tennison J.* The rdf data cube vocabulary. W3C candidate Recommendation. — 2013.— http://www.w3.org/TR/2013/CR-vocab-data-cube-20130625.

5. *Data Catalog Vocabulary (DCAT).*—2014. [2014 Feb 20].— http://www.w3.org/TR/vocab-dcat/

6. *Ismail M. A., Kareem S. A.* Identifying how novice researchers search, locate, choose and use web resources at the early stage of research// Malaysian Journal of Library and Information Science. — 2011. — Vol.16, No.3. — http://works.bepress.com/maizatulakmar_ismail/3.

7. *Calva-González J. J.* Las necesidades de información de los investigadores del área de Humanidades y Ciencias Sociales// Revista general de información y documentación.— 2003.— Vol. 13, No.2.— P. 155–180. — http://revistas.ucm.es/index.php/RGID/article/view/10799.

8. *Calva González J. J.* Las necesidades de información del usuario en la automatización de unidades de información// Revista Biblioteca Universitaria.— 2009.— Vol. 1, No.1.— P. 6. — http://www.dgbiblio.unam.mx/servicios/dgb/publicdgb/ bole/fulltext/volI1/necinf.html

9. *Sandusky R. J., Tenopir C.* Finding and using journal-article components: Impacts of disaggregation on teaching and research practice// Journal of the American Society for Information Science and Technology.— 2008.—Vol. 59, No.6.— P. 970–982. — doi:10.1002/asi.20804

10. *Bishop A. P.* Digital libraries and knowledge disaggregation: The use of journal article components// Proceedings of the third ACM conference on Digital libraries (p. 29–39). — 1998.— http://dl.acm.org/ citation.cfm?id=276679.— doi:10.1145/276675.276679

11. *Bishop A. P.* Document structure and digital libraries: How researchers mobilize information in journal articles// Information Processing & Management. — 1999.— Vol. 35,

No. 3.— P. 255–279.— doi: http://dx.doi.org/10.1016/j.bbr.2011.03.031

12. *Otlet P.* The science of bibliography and documentation// Rayward W. B (Org.). International organisation and dissemination of knowledge: Selected essays of Paul Otlet. — Amsterdam: Elsevier for the International Federation of Documentation, 1990. — https://www.ideals.illinois.edu/handle/2142/4004

13. *Rayward W. B. (Org.).* International organisation and dissemination of knowledge: Selected essays of Paul Otlet. (pp. 7-10). — Amsterdam: Elsevier for the International Federation of Documentation, 1990. Introduction. — https://www.ideals.illinois.edu/handle/2142/4004.

14. *Hjørland B., Nielsen L. K., Williams M. E.* Subject access points in electronic retrieval// Annual Review of Information Science and Technology. — 2001.—Vol. 35. — P. 249–298.

15. *Bates M. J.* The getty end-user online searching project in the humanities: Report No. 6: Overview and Conclusions// College & Research Libraries. —1996. — Vol. 57, No.6.— P. 514–523.— doi :10.1108/eb024508

16. *Markey K.* Twenty-five years of end-user searching, Part 1: Research findings// Journal of the American Society for Information Science and Technology. —2007.— Vol. 58, No.8.—P. 1071–1081.— doi : 10.1002/asi.20462.

17. *Markey K.* Twenty-five years of end-user searching, Part 2: Future research directions// Journal of the American Society for Information Science and Technology. — 2007.— Vol. 58, No.8.— P. 1123–1130. — doi:10.1002/asi.20601

18. *Rowlands I.* Electronic journals and user behavior: A review of recent research// Library & Information Science Research. — 2007.— Vol. 29, No.3.— P. 369–396.— doi : 10.1016/j.lisr.2007.03.005.

19. *Dogan R. I., Murray G. C., Névéol A., Lu Z.* Understanding PubMed® user search behavior through log analysis// Database: the Journal of Biological Databases and Curation.— 2009. — http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2797455/. — doi : 10.1093/database/bap018.

20. *Borgman C. L.* Why are online catalogs hard to use? Lessons learned from information retrieval studies// Journal of the American Society for Information Science. — 1986. — Vol. 37, No. 6.— P. 387–400. — doi : 10.1002/(SICI)1097-4571(198611)37:6<387::AID-ASI3>3.0.CO;2-8.

21. *Bates M. J., Wilde D. N., Siegfried S.* An analysis of search terminology used by humanities scholars: The Getty Online Searching Project Report Number 1// The Library Quarterly. — 1993.— P.1–39. — http://www.jstor.org/stable/4308771.

22. *Crowston K., Kwasnik B. H.* Can document-genre metadata improve information access to large digital collections// Library Trends. — 2003.— Vol.52, No.2.— P. 345–361. — http://works.bepress.com/cgi/viewcontent.cgi?article=1003&context=barbara_kwasnik.

23. *Shotton D.* Semantic publishing: The coming revolution in scientific journal publishing// Learned Publishing. — 2009. — Vol. 22, No.2. — P. 85–94. — https://webvpn.uminho.pt/http/0/www.ingentaconnect.com/content/alpsp/lp/2009/00000022/00000002/art00002. — doi:10.1087/2009202

24. *Scientific Data.* — 2014. — http:// www.nature.com/scientificdata/.

25. *Marcondes C. H., Mendonça M. A. R., Malheiros L. R., Da Costa L. C., Santos T. C. P.* Ontological and conceptual bases for a scientific knowledge model in biomedical articles// RECIIS.—2009.— Vol. 3, No.1. —http://www.reciis.cict.fiocruz.br/index.php/reciis/article/view/240/251. — doi:10.3395/reciis.v3i1.240en.

26. *Lee J. H., Downie J. S.* Survey of music information needs, uses, and seeking behaviors: Preliminary Findings// ISMIR. — 2004.— Vol. 2004, P. 5.— http://people.lis.illinois.edu/~jdownie/ismir2004_survey_downie_draft.pdf.

27. *Menzel H.* The information needs of current scientific research// The Library Quarterly 34.— 1964. — No. 1. — P. 4–19. — http://www.jstor.org/stable/10.2307/4305417

28. *Courtright C.* Context in information behavior research// Annual review of information science and technology. — 2007.— Vol. 41, No.1.— P. 273–306. — http://onlinelibrary.wiley.com/doi/10.1002/aris.2007.1440410113/full. — doi : 10.1002/aris.2007.1440410113.

29. *Berghel H., Berleant D., Foy T., McGuire M.* Cyberbrowsing: Information customization on the Web// Journal of the American Society for Information Science. — 1999.— Vol. 50, No.6.— P. 505–513. — http:// onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-4571(1999)50:6<505::AID-ASI5>3.0.CO;2-R/abstract.— doi:10.1002/ (SICI)1097-4571 (1999)50:6<505::AID-ASI5>3.0.CO;2-R.

30. *Tenopir C., King D. W., Edwards S., Wu L.* Electronic journals and changes in scholarly article seeking and reading patterns// Aslib proceedings. —2009.—Vol. 61.— P. 5–32. — http://www.emeraldinsight.com/journals.htm?articleid=1766871&show=abstract

31. *Tenopir C., King D. W., Spencer J., Wu L.* Variations in article seeking and reading patterns of academics: What makes a difference?// Library & Information Science Research.— 2009.— Vol. 31, No. 3.— P. 139-148.— http://www.sciencedirect.com/science/article/pii/S0740818809000516.

Как библиотеки и другие научные учреждения могут способствовать открытому доступу данных?*

Питер ЛИНДЕ
(Peter LINDE)

Технологический институт Блекинге,
Лен Блекинге, г. Карльскруна,
Швеция

Бриджит А. УЕССЕЛС
(Bridgette A. WESSELS),

Тордис СВЕЙНСДОТТИР
(Thordis SVEINSDOTTIR)

Шеффилдский университет,
Великобритания

Мерел НОРМАН
(Merel NOORMAN)

Голландская королевская академия искусств
и наук, Нидерланды

Изучаются вопросы, в чем состоит значение открытого доступа к научным данным и где и как библиотеки и связанные с ним заинтересованные лица могут способствовать получению выгоды от свободного обмена данными. В частности, акцентируется, как библиотекам стать компетентными при сотрудничестве по обучению и поощрению ученых и библиотечного персонала в сфере работы с открытыми данными. Статья основывается на ранних результатах проекта RECODE (EU FP7), изучающего стимулы и барьеры в развитии открытого доступа к научным данным в Европе (http://www.recodeproject.eu).

ВВЕДЕНИЕ

На протяжении последних 30 лет библиотеки приспосабливались к новым требованиям, тогда как аналоговые медиа превращались в цифровые. Библиотеки творчески подходили к преходящим увлечениям и/или долговременным реалиям, таким как Archie, Gopher, NCSA Mosaic, FTP, SGML, XML, Open Access, PDA и т.д. Сегодня большинство академических библиотек имеют учрежденческие архивы и цифровые издательские отделы, призванные поддерживать потребности ученых в распространении, сохранении документов и консультировании в открытом доступе. Библиотеки имеют большой опыт пропагандирования, обучения навыкам использования и внедрения публикаций открытого доступа, а также работы с цифровой информацией, но сегодня, когда мы наконец говорим о переломном моменте для научных документов открытого доступа [1], возникает новая «горячая» тема с полным современным набором требований к квалификации библиотек, бюджетам и организации – открытые данные*.

Открытый доступ к научным данным все больше рассматривается как позитивное развитие, поддерживаемое и стимулируемое внутри европейского научного ландшафта. Европейская комиссия продвигает научные данные к большей открытости в своей рамочной программе «Горизонт 2020»**, и эта тенденция также разви-

* Перевод Linde P., Wessels A. B., Sveinsdottir T., Noorman M. How can libraries and other academic institutions engage in making data open? //ELPUB 2014. Let's put data to use: Digital scholarship for the next generation, 18 th International conference on Electronic Publishing, 19-20 June, Thessaloniki, Greece.— http://elpub.scix.net/data/works/att/101_elpub2014.content.pdf

* К «открытым данным» мы относим научные данные, определяемые как любой материал, используемый в качестве основы для исследования.

** Пресс-релизы базы данных. Комиссия запускает пилотный проект по открытию финансируемых государством научных данных. 2013 г. (http://europa.eu/rapid/press-release_IP-13-1257_en.htm.)

вается внутри отдельных стран - членов ЕС и всего научного сообщества. Сегодня некоторые влиятельные журналы поощряют или требуют от ученых создания данных, которые бы поддерживали их публикации свободно доступными (например, Biomed Central journals, The Open Access Geoscience Data Journal Dataset Papers in Science, eLIFE, F1000Research и т.д.), тогда как национальные и частные финансовые организации причисляют открытый доступ к научным данным в качестве условия для финансирования. Однако получение открытого доступа и реализация его пользы требует значительной работы, как показывает растущий поток литературы по обмену данными и открытому доступу.

Сегодня кажется имеется больше общего согласия относительно значимости, которую открытые данные могут принести науке и обществу. По мнению их сторонников, неограниченный и облегченный цифровыми технологиями доступ к данным способен ускорить научный прогресс посредством минимизации дублирования усилий и предложения ученым более широкого ряда данных для использования в повторном анализе, сравнении, обобщении и проверке. Это принесет вклад в качество и целостность научных практик, так как увеличит прозрачность и ответственность. Также улучшится способ использования науки и научных данных в отношении социальных целей, а значит усилится значимость вносимого наукой в общество вклада. Более того, существует убежденность, что открытые данные будут выгодны для инновации и экономического роста. Европейская комиссия, например, относится к открытым данным как к «механизму для инноваций, роста и прозрачного управления» [2].

Но открытый доступ и повторное использование данных стали проблемой в большинстве научных дисциплин. Многие архивы, созданные для стимуляции обмена данными, остаются по большей части пустыми [3]. Несмотря на эти трудности, ряд передовых библиотек ощутил потребность в том, чтобы поддержать ученых в управлении и распространении научных данных. Мы более пристально рассмотрим некоторые из этих инициатив, часто начинающихся в виде проектов «новых возможностей», нацеленных на расширение библиотечных услуг в условиях, когда классические виды деятельности академических библиотек, такие как каталогизация, комплектование, подписные услуги и т.д., ставятся под сомнение и заменяются или автоматизируются. Существует много барьеров для открытия научных данных, и было бы не реально думать, что одно заинтересованное лицо в состоянии справится со всеми проблемами в одиночку. Существует острая потребность в кооперации изнутри, а также между организациями, разделяющая знание экспертов и специалистов.

Основным вопросом, выдвинутым в этой статье, является то, как библиотеки могут реализовать эту новую услугу вместе с другими заинтересованными в открытых данных лицами в научном мире?

В статье представлен обзор документов по научной политике, отчетов, научной литературы и других родственных документов, чтобы дать общее представление о текущих разработках в рамках данной области. Мы предоставляем анализ некоторых из этих подходов в целях определения надежных практик и вероятных барьеров*.

В современном, высоко конкурентном университетском климате продуктивность и качество являются лишь высокопарными словами, а финансирование научных исследований все в большей степени основывается на библиометрии. В такой среде для университетского управления становится более важным продолжать отслеживать продуктивность и качество научных публикаций. В то же самое время многочисленные финансовые фонды выступают за открытый доступ, а университеты борются за продвижение своей марки с целью заполучить лучших ученых и привлечь талантливых студентов.

На фоне такого ландшафта многие библиотекари осознают, что их услуги, включая архивы, одни из многих, которые должны взаимодействовать с целью поддержки и создания более обозримых научных данных.

Сегодня акамедические библиотеки изучают возможности интеграции учрежденческих архивов с системами CRIS (Current Research Information System – Информационная система по текущим научным исследованиям), работающих, как правило, с помощью университетских научных кафедр или аналогичных отделений [5]. В Швеции это изучается на национальном уровне, где национальный портал хранилищ SwePub вероятно будет объединен с системой CRIS шведских научных советов по научно-информационным системам*. Такие университеты, как Эдинбургский университет, объединили все научные услуги в одно отделение (информационные услуги), которое включает классические библиотечные функции, а также имеет подразделения типа: ИТ-инфраструктуры, цифровой центр курирования, национальный центр данных, разработанный объединенным комитетом по информационным системам (EDINA) и библиотека данных [6].

В своей «дорожной» карте изучения научных данных Лига европейских исследовательских университетов признала библиотеку в качестве основного источника управления и обнаружения данных [7]. Становится очевидным, что новая важная роль библиотеки, следующей дорогой электронной науки, состоит в том, чтобы являться компетентным командным игроком, когда дело касается построения таких структур поддержки для ученых. Наилучшим образом это осуществляется вместе с другими важными игроками в университете – службами научных кафедр, сотрудниками архива и служб научных ИТ и, безусловно, специалистами центра данных.

ПОТРЕБНОСТЬ В ОБУЧЕНИИ И ПРОПАГАНДЕ

Большинство ученых и вспомогательного персонала университета не знакомы с задачей управления открытыми данными, которая подразумевает огромную работу по пропаганде и обучению. В проекте «Возможности для обмена данными» (Opportunities for Data Exchange – ODE) [8] это выражено следующими словами: «Улучшение навыков и понимания учеными управления данными является необходимым. Подготовка должна начинаться в учреждениях, которые обучают ученых, на стадии аспирантур и выше, а возможно даже и раньше». Это часто указывает на то, что наилучшая практика сконцетрированного на дисциплине образования в сфере управления данными должна внедряться в обучение студентов и ученых на ранней стадии. Поэтому, чтобы играть активную роль в создании библиотек от-

* Данная статья основывается на наблюдениях, сделанных в ходе продолжающейся рабочих пакетов проекта RECODE [4].

* System för analys av svensk forskning.— http://www.mynewsdesk.com/se/kungliga_biblioteket/pressreleases/system-foer-analys-av-svenskforskning- 947591. Visited 140125.

крытых данных и обеспечении соответствующей компетентности, важным является как взаимодействие с другими университетскими заинтересованными лицами, так и активность в пропаганде и обучении в вопросах управления открытыми данными.

Одной из причин, почему обмен данными и открытый доступ до сих пор не являются нормой в большинстве дисциплин, служит то, что ученые сопротивляются тому, чтобы делать свои данные публичными. Их беспокойства лежат в следующих пределах, начиная от работы, которая будет взята из открытого доступа или неправильно использована, нехватки свободного времени или финансирования, чтобы сделать их данные доступными, и до сохранения неприкосновенности и конфиденциальности участников [3]. Также у ученых может отсутствовать опыт обмена данными [9]. Ученые выражают большое беспокойство относительно «объема работы и времени, необходимого для создания значимых и полезных данных в том случае, если они будут в открытом доступе. Например, время, затрачиваемое на аннотирование, создание и применение метаданных и контекста документа. Эта дополнительная работа потребует высвобождения времени из других видов научной деятельности, таких как сбор данных, анализ, публикации и обращения за финансированием, каждый из которых приносит явные и очевидные вознаграждения и выгоды для ученых и их карьеры [10]. Другая основная проблема заключается в том, что требуются значительные технические навыки по переводу данных в машиночитаемые форматы и использованию средств программного обеспечения для доступа и анализа данных. Ученые, желающие сделать свои данные публично и в цифровом виде доступными и повторно используемыми, должны ознакомиться со средствами программного обеспечения и форматами данных, которые не всегда легко могут подходить их существующим практикам исследования. В свою очередь, повторное использование данных потребует от ученых знания того, как искать и использовать данные с помощью сетевых средств. Также будет трудно найти общие стандарты и форматы для обмена данными, такие, которые другие смогут легко интерпретировать и использовать. Эти касающиеся практики барьеры также отражены в исследовании Европейской комиссии «Онлайн исследование по научной информации в цифровом веке» [11]. Почти 90% респондентов в этом исследовании не были согласны с утверждением: «Вообще говоря, в Европе НЕТ проблемы доступа к научным данным». Обеспечение подготовки ученых и технического персонала, а также создание осведомленности относительно возможностей и ограничений обмена данными будет поэтому способствовать превращению научных данных в более открыто доступные в различных дисциплинах.

Академические учреждения призваны играть ведущую роль в обучении и пропаганде. Исследование Комиссии также включило вопрос о том, как Европейский Союз может наилучшим образом содействовать доступу и сохранности научных публикаций и данных. Большинство респондентов уверенно согласилось с формулировками относительно «поддержания разработок европейской сети архивов» и «стимулирования университетов/научных учреждений, библиотек и финансовых органов и т.д. для осуществления определенной работы» [11]. Так как многие финансирующие органы уже возлагают ответственность на стратегии управления данными и согласования с научными учреждениями, то это также усиливает давление на научные учреждения с целью создания открыто доступных данных.

В самом научном сообществе наблюдается нехватка профессиональной подготовки в управлении данными, и реально никто не берет ответственность за функцию управления научными данными. В большинстве случаев библиотеки находятся в хорошем положении, чтобы взять на себя такую ответственность, но стандартный учебный план библиотечных школ не готовит студентов к управлению данными. Это необходимо менять.

РАЗЛИЧНЫЕ КУЛЬТУРЫ И ЦЕЛЕВЫЕ ГРУППЫ

В изученном материале прослеживается общее наблюдение, что ученые представляют собой очень неоднородную группу. Не только по дисциплинам, но и между собой в рамках одной и той же команды. Поэтому важно получить понимание «культуры» внутри любого коллектива ученых, прежде чем рассматривать, как влиять на их поведение в отношении управления научными данными [10].

Научные данные отличаются от публикаций. Они более разнообразны и часто связаны с проектами научных сообществ, требующими новых способов работы, мышления и взаимодействия для библиотекарей. Разнообразие данных, инструментов и потребностей ученого должны измеряться не на уровне дисциплины, а на уровне научной группы.

Рекомендуется, чтобы в целях пропаганды и обучения интервью, случайные и научные исследования разрабатывались так, чтобы понять научные требования и поведение [12,13,16,17,19]. Это должно стать основой для разрабатывающихся материалов по пропаганде/обучению, что будет мотивировать ученых, а также заставит их понимать обязательства перед финансовыми организациями, учреждениями и общественностью. Подготовка планов по управлению данными и обучение персонала для их реализации являются новым и малоизученным делом для университетов и научных учреждений, но имеются хорошие примеры этого и отчеты о том, как поддержать эти организации в управлении открытыми данными.

Марк Л. Браун и Уэнди Уайт рассказывают историю о том, как Саутгемптонский университет через сотрудничество со Службой научных данных Соединенного Королевства и вовлечение в такие проекты как IDMB начал улучшать и формализовать инициативы, чтобы поддержать ученых университета в управлении их научными данными [14].

Что касается целей обучения, то было начато использование автоматизированных и сетевых средств. Например, автоматизированные средства для поддержки создания идентификаторов цифровых объектов (DOI) DataCite и руководства на основе сети, чтобы помочь интерпретировать требования финансовых сторон.

Что касается услуги для ученых по планированию управления данными, то была разработана программа обучения с целью вовлечь различные группы, начиная от аспирантов до научных сотрудников. Планирование и реализация этих курсов, лекций, практикумов и семинаров всегда проводились совместно с учеными.

В консультационном отчете, сделанном для объединенного комитета по информационным системам [15], изучались роли, права, распределение ответственности и отношения организаций, центров данных и других заинтересованных лиц, работающих с данными. Выводы относительно пропаганды и обучения весьма схожи с выводами из Саутгемптона: важность целевых и адаптационных мер для отдельных дисциплин и поддисциплин; осознание курирования данных и сохранения хо-

роших практик является в целом низким, но значительно варьируется между дисциплинами; рекомендации для центра данных и персонала учрежденческого архива выходить к пользователям и продвигать свои обучающие программы с помощью сочетания методов, семинаров, практикумов, уроков и т.д.

Как сообщается в большей части литературы, важной целевой группой для пропаганды и обучения по управлению открытыми данными являются молодые ученые и студенты старших курсов и далее. В первую очередь пропаганда должна касаться сообщества выпускников и аспирантов, так как они находятся на переднем крае как сборщики и генераторы данных и, конечно, как будущие ученые [16].

СНИЗУ ВВЕРХ ИЛИ СВЕРХУ ВНИЗ?

Типичной американской программой курирования данных является «свобода от мандатов и побуждений со стороны верхнего уровня, но обогащение посредством независимого действия «снизу вверх»». Подобная ей структура основана на предприимчивости индивидуумов и сопровождается медленным темпом развития [17]. В недавнем американском исследовании, направленном на определение современных тенденций в управлении научными данными в научных организациях, только 9% респондентов утвердительно ответили на вопрос: «Есть ли у вашей организации политика управления данными?». Почти 90% согласилось со следующим утверждением – «Политика управления данными на уровне учреждения является важной», это показывает, что заинтересованные в рамках университета, стороны, такие как ученые, библиотекари, администрация научного персонала, преподаватели и т.д., очень хотят увидеть внедрение таких стратегий [22].

Причина, по которой библиотеки вообще начали осуществлять программы курирования данных, обосновывается их передовой позицией относительно публикаций архивов открытого доступа и инициативами цифрового хранения, ранее изученными университетскими библиотеками. Также утверждается, что это даст библиотеке возможность стимулировать существующие партнерские отношения и вовлечь их в новые с целью выработки навыков и необходимых связей для курирования данных. Объединение с несколькими научными сообществами в качестве отправной точки является способом получить одобрение, форму и программные обязательства со стороны администрации. Успешный проект может быть достаточным поводом убедить университетских администраторов в пользе политики курирования на уровне университета и мандата [17].

В Саутгемптоне [15] реакция на то, что финансовые стороны все в большей степени возлагали ответственность за стратегии управления данными и согласование с научными учреждениями, вызвала появление подхода «снизу вверх», основанного на потребностях ученых, и желание разработать требования для подхода «сверху вниз» и соответствующей инфраструктуры. Их опыт с хранилищами издательств открытого доступа показал, что «ученые открыты новой практике пока ими руководят, интегрируют в научный поток, отражающий дисциплинарные различия, и поддерживают консультациями и обучением. Ясность в отношении политики и соответствующей службы поддержки необходима».

Весьма важным было то, что организации в составе университета понимали: они распоряжаются инвестициями и службами поддержки, касающимися управления данными, не испытывая принуждения со стороны многочисленных требований.

В этом процессе окончательная политика управления данными возлагала ответственность за запись, поддержание, хранение, безопасность и т.д., при соблюдении определенных правил, на ученых, что хорошо с точки зрения библиотеки. Разумно и то, что создатели данных также записывают их, а библиотека является здесь поддерживающей, а не принуждающей стороной этого процесса.

Исходя из основных компонентов Саутгемптонского проекта, структура учрежденческой политики, рабочий регистр данных, консультирование по управлению данными и руководства, а также соответствующая бизнес-модель – это университетская политика по управлению научными данными, которая считается наиболее важной. В конце концов и по причине существования баланса сил имеется потребность в формальном мандате или политике, исходящей от более высокой университетской власти [14].

Библиотекари ввели в действие и управляют учрежденческим хранилищем и идеей открытого доступа с большим знанием в вопросах научной коммуникации, но поскольку они не приносят никаких финансовых средств университету, библиотека в большинстве случаев воспринимается как обслуживающее подразделение, но без особого влияния. Между тем она (библиотека) рассматривается в качестве первого шага к формальной политике, когда нет четкого руководства со стороны правительства, а администрация университетов воздерживается от решений, касающихся ресурсов или инициатив по вопросам управления данными, подход «снизу вверх» представляет собой стартовый момент, где пропаганда является только первым шагом.

НОВЫЕ РОЛИ И ПАРТНЕРЫ

Саутгемптонский университет является одним из примеров того, как инициативы проектов по курированию данных не перестают работать при сотрудничестве внутри отделений университета. На протяжении долгого времени необходимые навыки доступны только через партнерство с внешними учреждениями или организациями [17,18].

Независимо от того, как библиотеки подходят к проблеме курирования данных, введение новых навыков в библиотечную профессию крайне необходимо. Работа библиотекарей с учеными, в которой они будут выполнять роль «библиотекарей данных», должна содержать навыки как технической, так и архивной работы в отношении данных (подобно двум сторонам монеты). Такого рода специалисты будут играть основную роль в процессе научного издательства и должны соответствующим образом поощряться. Библиотечным школам необходимо ввести курсы, отвечающие этим новым должностным инструкциям.

Имеется безусловная потребность в слиянии библиотечных и архивных навыков с целью сделать университетские хранилища хорошо функционирующим местом для открытых данных. Также это может быть частью профессионального роста и обучения [14]. Справедливо это и для библиотечных специалистов, что противопоставляется бюро по изучению специалистов, которые находятся ближе к ученым, поддерживающим их посредством внедрения проектов, статистических данных и т.д. Это может также быть шансом для ролей классических библиотек, например, как слияние библиотекарей с целью дальнейшего расширения. Слияния

могут помочь ученым депонировать данные на момент их создания. Они могут посоветоваться относительно стандартов, применяемых к потребностям, создать планы курирования для целого жизненного цикла данных в полном соответствии с мандатами финансовых сторон [19].

Как утверждалось ранее, уровни навыков ученых относительно управления данными разнообразны, и обучение здесь более необходимо. Поэтому параллельно пропаганде существует требование развития навыков сообщества. Но так как большая часть экспертизы в управлении данными сконцентрирована в центрах данных, то существует потребность привлечения и формализации потока знания из центров данных в учреждения, в которых сотрудники сегодня все больше назначаются на управление и разработку хранилищ для курирования данных.

С 1976 г. Консорциум европейских архивов данных по общественным наукам служил в качестве неформальной головной организации для европейских национальных архивов данных. Архивы данных Консорциума и другие подобные тематические архивы находятся в хорошей позиции, чтобы работать с университетскими библиотеками и вести переговоры с архивами по поводу обучения.

Иногда наблюдается поляризация точек зрения относительно роли учрежденческих хранилищ для данных. Центры и архивы данных имеют более длительную перспективу, чем учрежденческие хранилища, являющиеся относительно новыми структурами и еще только доказывающими свои возможности. Но и центры данных, и библиотеки играют руководящую роль в деятельности по курированию данных. И те, и другие помогают и направляют ученых на хранение своих данных. Разделение ролей – краткосрочное, легкодоступное хранение, осуществляемое усилиями учрежденческих хранилищ, и долговременное хранение, реализуемое центрами данных, может быть одним из способов, облегчающим поддержку относительно управления данными и взаимодействия [14].

ВЫВОДЫ И ОБСУЖДЕНИЕ

Вопрос о новых ролях библиотек в управлении открытыми данными безусловно является одним из вопросов о финансировании новых услуг. Очевидна необходимость в том, чтобы университет создал экономический план расходов на хранение, курирование, обучение и т.д. в отношении научных данных.

Возможно основной проблемой будет убедить администрацию университета в сборе экономических ресурсов для разработки моделей курирования данных. Действительно большая часть недостающего финансирования для управления научными данными происходит из самих библиотек [18]. Как правило, у организации нет в наличии дополнительных денежных средств, и библиотеки вынуждены либо перераспределять внутренние ресурсы, либо находить внешнее финансирование, например путем взаимодействия с внешними партнерами. Поэтому введение системы грантов и финансирование библиотек на национальном и международном уровнях будет важным фактором, заставившим ускорить курирование данных на более широком уровне в университетах [17].

По-видимому, не будет реального роста финансирования без учрежденческих или национальных мандатов, способствующих внедрению планов управления научными данными. Практики снизу вверх являются медленными генераторами изменений и общей согласованности и должны будут дополняться формальными политиками.

Среди основных научных заинтересованных сторон в экосистеме открытых данных мы имеем финансирующие науку стороны – советы и фонды; создателей данных – ученых, а также распространителей и кураторов данных, в этом случае это библиотеки, архивы и центры данных. Все эти заинтересованные стороны с их организациями будут нужны для взаимодействия, так как барьеры многочисленны и сложны, и только совместные силы смогут реализовать идею открытых данных. Финансисты и разработчики политики нужны, чтобы четко курировать управление данными, а также выделять средства на обучение, инфраструктуру, проекты курирования данных и т.д. Профессиональные ассоциации должны побуждать к созданию новых возможностей для обучения специалистов. Библиотекари, специалисты в области информационных технологий, штат научных сотрудников из университетов нуждаются в сотрудничестве с архивистами и кураторами центров данных и наоборот. Ученым следует установить новые приоритеты относительно важности управления данными, найти способы самоокупаемости.

Такое сотрудничество уже ведется, но должно расширяться и стимулироваться правительственными и научными властями, которые проводят политику, направленную на облегчение сотрудничества и прояснение дорожных карт для будущей работы. Одинаковую важность представляют неправительственные группы по пропаганде и другие межпрофессиональные организации, заинтересованные в продвижении вперед вопроса об открытых данных. Такие организации, как COAR, EUDAT, LIBER [20], RDA, SHERPA, SPARC, KE и многие другие проделывают фантастическую работу по пропаганде и информированию о важности управления данными и служат неисчерпаемым ресурсом для библиотек, которые желают начать реализацию схем курирования данных.

В настоящее время существует временный пробел между техническим знанием и доступом к соответствующей инфраструктуре, но помимо этого среди библиотек и библиотекарей нет понимания сложности процесса управления открытыми данными. Опираясь на опыт отдельных исследований, выполненных для проекта RECODE, мы утверждаем, что значимость неограниченного доступа к научным данным преимущественно зависит от качества процесса открытого доступа. Наш анализ ценностей и мотиваций среди ученых относительно открытого доступа показал, что подходы к поддержке и совершенствованию развития открытого доступа к научным данным нуждаются в изучении, по крайней мере, следующих вопросов:

- Они (подходы) должны быть чувствительны к различным научным практикам для обеспечения гарантии того, что существующая научная точность поддерживается наравне с облегчением открытого доступа.
- Они должны установить связь между инфраструктурами, юридическими и этическими вопросами и учрежденческими подходами таким образом, чтобы экосистема открытого доступа могла поддерживать соответствующий подход ко всем типам данных в рамках их научных областей.
- Они нужны, чтобы обеспечивать сохранность анонимности и конфиденциальности участников.
- Они должны обеспечить способы для цитирования и правильного отражения принадлежности открытых данных, чтобы стать частью этической научной практики.

- Им нужно обратить внимание на технические вопросы, например способы, с помощью которых технология управляет огромным массивом данных; отсутствие технической инфраструктуры для решения вопросов хранения данных и взаимодействия.
- Культурные барьеры имеют существеннное значение, особенно такие вопросы, как конкуренция за награды и репутацию в науке, отсутствие доверия между учеными и отсутствие карьерного роста, связанного с наградами и престижем, являющимися результатом публикации и обмена данными.

Для бибилиотек жизненно важно осознать, что сегодня пришло время быть сторонниками вовлечения в управление научными данными – введение профессиональных программ подготовки, разработка пилотных программ, мониторинг основных инициатив данных, таких как DataCite, DataONE и т.д. и хороших примеров библиотечных инициатив Эдинбургского университета [6], Университета Йорка*. Саутгемптонский университет, центр курирования распределенных данных университета Пурдью** или другие рискуют быть обойденными другими участниками на арене установления программ управления научными данными. Роль библиотек в обучении управлению данными очевидна не всем. Некоторые ученые согласны, что бибилиотекам следует иметь и играть все более важную роль управляющих данными и экспертов на основе их положения в издании статей открытого доступа. Другие ученые утверждают, что центры данных могли бы обеспечить необходимую поддержку для правильной обработки данных [8]. Сейчас самое лучшее время для освещения этих вопросов и начала изучения экспериментов, проведенных на данный момент в целях выполнения важной задачи сделать научные данные открытыми. Если библиотека не видит потенциала в задаче относительно новаторства в сфере открытых научных данных, как это имеет место в случае пропаганды открытого доступа к научным публикациям, то существует огромный риск, что другие заинтересованные стороны быстро освоят эту роль и расширят услуги по отношению к ученым, а библиотекари останутся без ответа на вопрос, – как библиотеки могут участвовать в процессе создания открытого доступа данных.

ЛИТЕРАТУРА

1. *Archambault E. et al.* Proportion of open access peer-reviewed papers at the European and World levels – 20014-2011. August 2013. Produced for the European Commission DG Research & Innovation by Science-Metrix Inc.

2. *European Commission.* Open Data, an engine for innovation, growth and transparent governance, COM 882 final, Brussels, 12 December 2011. — 2011.— http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF

3. *Nelson B.* Data sharing: Empty archives// Nature. — 2009.— Vol. 461.— P. 160-163.

4. Policy RECommendations for Open access to research Data in Europe. — http://recodeproject.eu/

5. *Joint N.* Current research information systems, open access repositories and libraries// Library Review. — 2008.— Vol. 57, No.8.

6. *Rice R. et al.* Implementing the research data management policy: University of Edinburgh Roadmap// International Journal of Digital Curation. — 2013.—Vol. 8, No.2.

7. LERU roadmap for Research Data. League of European Research Universities, 2013. — http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf

8. *Dallmeier-Tiessen S. et al.* Compilation of results on drivers and barriers and new opportunities. — 2012.— http://www.alliancepermanentaccess.org/wpcontent/uploads/downloads/2012/08/ODECompilationResultsDriversBarriersNewOpportunities1. pdf.

9. *Borgman C. L.* The Conundrum of sharing research data// Journal of the American Society for Information Science and Technology. — 2012. — Vol 63, No.6.

10. *Sveinsdottir T. et al.* Deliverable D1: Stakeholder values and ecosystems. Policy RECommednations for Open access to research Data in Europe (RECODE), 30 september 2013.— http:// recodeproject.eu/wp-content/uploads/ 2013/ 10/ RECODE_D1-Stakeholder-values-andecosystems_Sept2013.pdf

11. *European Commission.* Online survey on scientific information in the digital age.— 2012. ISBN: 978-92-79-23170-4. DOI:10.2777/7549

12. *Lyon L. et al.* Final report – Disciplinary approaches to sharing, curation, reuse and preservation. Jisc 2009. — http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf

13. *Schmidt L., Ghering C., Nicholson S.* Digital curation planning at Michigan State University// Notes on Operations. —2011. — Vol. 55, No2. — http://staff.lib.msu.edu/nicho147/Research/DigCur_LRTS_2011.pdf

14. *Pryor G., Jones S., Whyte A.* Delivering research data management services: Fundamentals of good practice. — Facet Publishing, 2013.

15. *Lyon L.* Dealing with data: Roles, rights, responsibilities and relationships. — Consultancy Report, 2007.

16. *Carlson J. R., Bracke M. S.* Data management and sharing from the perspective of graduate students: An examination of culture and practice at the Water Quality Field Station// Libraries Faculty and Staff Scholarship and Research.— 2013.— Paper 53.

17. *Walters T.* Data curation program development in U.S. Universitites: The Georgia Institute of Technology Example// The International Journal of Digital Curation. — 2009. — Vol. 4, No.3.

18. *Research Data Management* – Principles, practices, and prospects// Council on Library and Inforation Resources. — 2013. ISBN 978-1-932326-47-5. —http://www.clir.org/pubs/reports/pub160

19. *Gabridge T.* The last mile: Liason roles in curating science and engineering research data// Research Library. Issues: A bimonthly report from ARL CNL and SPARC, August 2009. http://old.arl.org/bmdoc/rli 265 gabridge.pdf

20. *Chrisensen-Dalsgaard et al.* Ten recommendations for libraries to get started with research data management. Final report of the LIBER working group on E-Science// Research Data Management, 2012.

* Research data management at the university of York. — http://www.york.ac.uk/about/departments/support-and-admin/informationdirectorate/strategy/projects/rdm/

** Distributed data curation center, D2C2. Purdue University Libraries. —http://d2c2.lib.purdue.edu/Purdue University

Публикуйте ваши данные и код модели: научный выход – это больше, чем «просто» научная статья*

Мартин РАСМУСЕН

(Martin RASMUSEN)

Издательство Copernicus Publications (Copernicus GmbH), г. Гёттинген, Германия

Научный выход (продукция) это не только научные статьи. В целях обеспечения рынка сбыта публикациями другого вида, не статьями, инновационное издательство открытого доступа Copernicus Publications, базирующееся в г. Гёттинген (Германия), выпускает журналы "Earth System Science Data" и "Geoscientific Model Development". Первый журнал посвящен рецензируемой публикации статей по множествам оригинальных научных данных в системе наук о Земле. Второй – публикует описание, развитие и оценку цифровых моделей земной системы и ее компонентов. Оба журнала применяют инновационный интерактивный доступ относительно рецензирования, с отчетами общественных рецензентов, публичными комментариями членов сообщества, имеющими место до решения редактора, и публичными откликами авторов. Мотивация состоит в том, чтобы сделать всю научную продукцию – от данных до моделей, до научных результатов и новых интерпретаций – свободно доступной, с целью поощрения научной дискуссии, увеличения прозрачности в гарантии научного качества и оказания доверия всем вовлеченным в процесс участникам

ВВЕДЕНИЕ

В науках о земной системе, как и во многих других дисциплинах, окончательная интерпретация новых научных сведений является результатом длительного процесса сбора данных, объяснения данных, выверки модели, работы модели, толкования этих результатов и выводов относительно аспектов новизны. Это групповая работа команды многих людей, вносящих вклад в эти результаты, сюда входят не только ученые, но и инженеры, специалисты в области данных и многие другие группы знающих участников.

Когда началось издательство открытого доступа, то быстро возникла идея распространить этот принципы на многие другие научные источники, а не просто на окончательную, исправленную, отрецензированную статью. В течение десятилетий читатели научных статей должны были решать вопросы относительно диаграмм, являющихся результатом интерпретации данных или работы модели, не зная многого о происхождении данных и структуре без получения доступа к этим данным, без широкого понимания используемых моделей и без глубокого осознания кода модели. Ни рецензенты, ни читатели не могли даже воспроизвести работу автора научной рукописи.

К счастью, принцип открытого доступа стал в плане политики широко принятой стратегией, и либеральное авторское право и лицензионные соглашения типа лицензии Creative Commons' CC-BY фундаментально переработали идею доступа к научной работе и опциям для повторного использования в большинстве случаев научной продукции, профинансированной через средства налогоплательщиков.

В 2008 г. две группы ученых, независимо друг от друга, выдвинули идею журнала публикации данных, с одной стороны, и журнала разработки модели, с другой. Издательство «Copernicus Publications» начало издавать эти два журнала, применяя подход издательства интерактивного открытого доступа с публичным рецензированием и интерактивной публичной дискуссией, установленный в 2001 г. Отчеты общественных рецензентов, публичные комментарии со стороны сообщества, предшествующие решению редактора, и публичные отклики авторов публикуются вместе с доступной для рецензирования версией авторской рукописи. Мотивация заключалась в том, чтобы сделать весь научный выход от данных до моделей, до научных результатов и новых интерпретаций свободно доступным, поощрить научную дискуссию, увеличить прозрачность в гарантии научного качества и вызвать доверие ко всем вовлеченным в процесс участникам.

• Перевод Rasmusen M. Publish your data and model code: Research output is more than "just" a research paper. – http://elpub.scix.net/data/work/att/115_elpub2014.content.pdf

Последующие разделы описывают журналы *Earth System Science Data* (ESSD) и *Geoscientific Model Development* (GMD) более подробно и объясняют концепцию издательства интерактивного открытого доступа.

ЖУРНАЛ *EARTH SYSTEM SCIENCE DATA* (ESSD)

Цели, область и мотивация

Earth System Science Data (ESSD) – международный, междисциплинарный журнал для публикации статей по оригинальным научным данным, способствующих повторному использованию высококачественных данных, полезных для системы наук о Земле. Редакторы поощряют представление материалов по оригинальным данным или массивам данных, являющихся сведениями достаточного качества и потенциального воздействия, чтобы внести вклад в осуществление целей журнала. Он содержит разделы для обычных статей регулярного объема, кратких сообщений (например, по дополнениям к наборам данных) и комментариев, а также обзорных статей и специальных выпусков.

Статьи в разделе данных могут иметь отношение к планированию, инструментарию и проведению экспериментов или к сбору данных. Любая интерпретация данных находится вне области регулярных статей. Методические статьи описывают нетривиальные статистические и другие применяемые методы, например, фильтрация, нормализация или конвертирование необработанных данных в первичные опубликованные данные, а также нетривиальный инструментарий или операционные методы. Любое сравнение с другими методами находится вне сферы регулярных статей. Обзорные статьи могут сравнивать методы или относительные достоинства наборов данных, пригодность отдельных методов или наборов данных для определенных целей или как комбинации методов могут быть использованы в качестве более сложных методов или массивов справочных данных.

Этот журнал ставит своей целью создать новый предмет публикации: публиковать данные в соответствии с общепринятой формой публикуемых статей, используя установленные принципы оценки качества через рецензирование и до анализа наборов данных. Задачи заключаются в том, чтобы сделать наборы данных надежным ресурсом, на котором можно основываться, и вознаграждать авторов с помощью установления приоритета и признания через влияние их статей.

Рецензирование коллегами гарантирует, что наборы данных являются, по крайней мере, внушающими доверие и не содержат никаких могущих быть обнаруженными проблем, что они достаточно высокого качества и их ограничения четко сформулированы, что они представляют собой открытый доступ (свободный от платы), хорошо аннотированы с помощью стандартных метаданных (например, ISO 19115) и доступны из сертифицированного центра данных/хранилища, и что они являются привычными в отношении их формата(ов) или протокола доступа, однако на них не распространяется право собственности (например, стандарты Open Geospatial Consortium standards), ожидается, что наборы данных станут пригодными для использования в предсказуемом будущем.

Статьи в этом журнале должны давать возможность рецензенту и читателю просматривать и соответственно использовать данные, затрачивая при этом меньше усилий. Поэтому вся необходимая информация должна быть представлена через текст статьи и ссылки в сжатой манере, а каждая статья должна публиковать по возможности как можно больше данных. Цель заключается в том, чтобы минимизировать всю рабочую нагрузку рецензентов, например, путем рецензирования одной вместо многих статей и максимизировать влияние каждой статьи [1].

Создателями ESSD были Дэвид Карлсон, директор программного бюро Международного полярного года (International Polar Year – IPY) в 2007-2008 гг., и Ганс Пфайфенбергер, руководитель инфраструктуры информационной технологии в Институте полярных и морских исследований Альфреда Вегенера (Alfred Wegener Institute for Polar and Marine Research – AWI) в Бремерхафене, Германия.

Представление рукописи

Предварительным условием представления рукописи для публикации в журнале ESSD является то, что наборы данных, на которые ссылаются в рукописи, должны быть представлены в долгосрочное хранилище. Такое хранилище должно отвечать следующим основным критериям при любых обстоятельствах [1]:

- Постоянный идентификатор: наборы данных должны иметь идентификатор цифрового объекта.
- Открытый доступ: наборы данных должны быть доступны бесплатно и без каких-либо препятствий, за исключением обычной регистрации для получения свободного входа в систему.
- Либеральное авторское право: любой может свободно копировать, распространять, передавать и адаптировать наборы данных пока он доверяет оригинальным авторам (эквивалент Creative Commons Attribution License).
- Долгосрочная доступность: хранилище должно отвечать наивысшим стандартам, чтобы гарантировать долгосрочную доступность наборов данных и постоянный доступ.

Критерии рецензии

Для будущего повторного использования и интерпретации пользователь обязательно должен быть уверен в качестве научных данных. Задача ESSD в том, чтобы обеспечить оценку качества наборов данных, которые уже находятся в постоянных хранилищах. Подходит ли сама статья для поддержки публикации набора данных? Является ли набор данных важным – уникальным, полезным и полным? Является ли публикация набора данных, в том виде как она представлена, высокого качества? Рецензентов просят принять решение относительно того, насколько хорошо наборы соответствующих данных, представленных статьей, и сама статья отвечают критериям значимости, качества данных и качества представления [1].

ESSD – факты и цифры

На конец марта 2014 г. ESSD имел 127 представленных рукописей, из которых 99 были опубликованы на дискуссионном форуме ESSD, а 85 – в самом ESSD в качестве окончательных и исправленных журнальных статей. Готовые статьи имеют средний объем в 12 страниц (медиана), а рецензирование в среднем (медиана) составляет 29 дней с момента представления дискуссионной статьи к публикации и 33 дня с момента представления исправленного варианта после публичной дискуссии до окончательной публикации исправленной

и полностью отрецензированной статьи. На дискуссионном форуме было получено 433 комментария, 207 из которых являются комментариями рецензента, 194 – комментариями автора, восемь комментариев опубликованы журнальными редакторами и 24 членами научного сообщества до окончательного принятия рукописей [2]. ESSD индексируется базой данных Scopus.

РАЗВИТИЕ ГЕОНАУЧНОЙ МОДЕЛИ

Цели, область и мотивация

"Geoscientific Model Development" – международный научный журнал, посвященный публикации и публичной дискуссии относительно описания, разработки и оценки цифровых моделей земной системы и ее компонентов. Типы рукописей, рассмотренные для рецензируемой публикации, следующие [3]:

- Описания геонаучной модели – от небольших моделей до GCM.
- Развитие и технические статьи, описывающие такое развитие как новые параметризации или технические аспекты действующих моделей, таких как воспроизводимость результатов.
- Статьи, описывающие эксперименты с новыми стандартами для оценки работы модели или новейшие способы сравнения результатов модели с данными наблюдения.
- Описания взаимного сравнения моделей, включая подробности эксперимента и протоколы проектов.

Журнал GMD принадлежит Европейскому союзу по геонаукам (European Geoscience Union – EGU; EGU, http://www/egu/eu) и начал издаваться с 2008 г. Основными проводниками идей журнала и исполнительными редакторами были (в алфавитном порядке): Джеймс Аннан и Джулия Харгривс, оба из Научно-исследовательского института глобальных изменений – JAMSTEC, г. Йокагама, Япония; Дэн Лант, Бристольский университет, Великобритания; Роберт Маш, Саутгемптонский университет, Великобритания; Энди Риджвел, Бристольский университет, Великобритания; Айан Рутт, Университет Суонси, Великобритания; Рольф Сандер, Химический институт Макса Планка, г. Майнц, Германия.

Поскольку масштаб и сложность средств компьютерного моделирования увеличились, то больше не применялась практика описания моделей в статьях. Кроме того, обычная журнальная рецензия концентрируется на научных результатах и описании модели, а технические детали представлены менее хорошо. Однако инициаторы GMD увидели потребности в том, чтобы достаточно полно описать модели и их разработки в рецензируемых публикациях. Они поставили своей целью гарантировать воспроизведение, тщательное прослеживание, прозрачность и доступ [4]. На сетевом сайте GMD приведены две следующие тщательно подобранные цитаты:

«Я полагаю, что настало время для значительно лучшей документации программ, и что мы можем наилучшим образом достичь этого, считая, что программы должны быть грамотными работами». (Donald E. Knuth. Literate Programming, 1984)

«По существу все модели ошибочны, но некоторые полезны». (George E.P. Box. Robustness in the strategy of scientific model building, 1979).

Критерии рецензии

Рецензентов просят определить научную значимость, научное качество, возможность научного воспроизводства, а также качество представления. Рецензенты решают, описываются ли существенно новые концепции, идеи или методы; обоснованы ли подходы и применяемые методы; имеют ли модели потенциал произвести вычисления, ведущие к важным научным результатам, и до какой степени наука моделирования является воспроизводимой. Члены научного сообщества должны быть способны воспроизводить науки. Поэтому центром внимания становится полнота и точность описаний [3].

GMD – факты и цифры

К концу марта 2014 г. журнал GMD имел 596 представленных рукописей, из которых 495 опубликованы на дискуссионном форуме GMD и 351 непосредственно в самом GMD в качестве готовых, отрецензированных журнальных статей. Средний объем готовых статей – 16 страниц (медиана), а рецензирование занимает в среднем (медиана) 33 дня с момента представления до публикации дискуссионной статьи и 46 дней с момента от представления исправленного варианта после публичной дискуссии до публикации окончательно исправленной и полностью отрецензированной статьи. На дискуссионный форум было отправлено 2 129 комментариев, из которых 1 018 являются комментариями рецензентов, 963 – авторскими комментариями, 67 комментариев опубликованы редакторами журналов и 81 – членами научного сообщества, все они были отправлены до окончательного принятия рукописей [5]. GMD индексируется в базах данных Scopus и Web of Science, журнал получил импакт фактор издательства Thomson Reuters со значением 5, 030 в 2012 г. [3].

ИЗДАТЕЛЬСТВО ИНТЕРАКТИВНОГО ОТКРЫТОГО ДОСТУПА

Издательство интерактивного открытого доступа предназначено для внесения большей прозрачности в гарантию научного качества с помощью свободно доступной публикации отчетов рецензентов и откликов автора. На первом этапе представленная рукопись рецензируется на предмет принятия одним из тематических редакторов журнала. Это представляет собой быстрый просмотр и внесение только технических исправлений. Затем рукопись набирается с помощью компьютера и публикуется в качестве так называемой дискуссионной статьи. Она является полностью цитируемой, получает классические ссылки и пагинацию, а также DOI. Публикационная платформа называется дискуссионным форумом.

Затем дискуссионная статья становится предметом интерактивной публичной дискуссии, во время которой комментарии рецензентов (анонимных или назначенных), дополнительные краткие комментарии других членов научного сообщества (назначенных) и ответы авторов также публикуются на дискуссионном форуме вместе с дискуссионной статьей. В отличие от других инициатив, экспериментирующих с публичным рецензированием, комментарии в этой концепции также являются полностью цитируемыми, пронумерованными, автоматически набранными с помощью использования онлайн, и остаются в режиме онлайн постоянно.

На втором этапе процесс рецензирования завершон и, в случае принятия, готовые (исправленные) отрецензированные статьи публикуются в самом журнале. Последний тогда служит полностью отрецензированной публикационной платформой, которая является предметом отражения в Web of Science, Scopus и других базах данных.

Концепция интерактивного издательства открытого доступа стартовала в 2001 г. и прослеживается до Ульриха Пёшля и Нобелевского лауреата Пауля Крутцена, оба из Химического института Макса Планка в Майнце (Германия). Впервые она была применена к журналу *Atmospheric Chemistry and Physics* (ACP) [6], очень успешному изданию, принадлежащему Европейскому союзу по геонаукам (EGU) и выпускаемому издательством Copernicus Publications.

Ульрих Пёшль описал свою концепцию во многих публикациях [7-9].

ЛИТЕРАТУРА

1. ESSD journal website at http://www.earth-system-data net. Access on 27 March 2014/

2. ESSD paper statistics, taken from Copernicus Publication' manuscript review system Copernicus Office Editor. Access on 27 March 2014.

3. GMD journal website, available at http:// www.geoscientific-model-development net/. Access on 27 March 2014.

4. GMD Executive Editors: Editorial: The publication of geoscientific model developments v1/0, Geosci. Model Dev., 6,1233-1242, doi:105194/gmd-6-1233-2013, 2013.

5. GMD paper statistics, taken from Copernicus Publication' manuscript review system Copernicus Office Editor. Access on 27 March 2014.

6. ACP journal website, available at http:// www.tmospheric-chemistry-and-physics.net/ Access on 27 March 2014.

7. Pöschl U. Interactive journal concept for improved scientific publishing and quality assurance// Learned Publishing. – 2004. – Vol. 17. – P.105-113.

8. Pöschl U. Interactive Open Access Publishing and Peer Review: The Effectiveness and perspectives of transparency and self-regulation in scientific communication and evaluation//Liber Quarterly. -2010. –Vol. 19. – P. 293-314.

9. Pöschl U. Multi-stage open peer review: Scientific evaluation integrating the strengths of traditional peer review with the virtues of transparency and self-regulation// Frontiers of Computational Neuroscience. -2012. – Vol. 6, No. 33. – doi: 10.3389/fncom.2012.00033, 2012.

Выполнение Рекомендации Европейской комиссии по открытому доступу к научной информации: сравнение национальных политик*

Лисиане ЛОМАЦЦИ
(Lisiane LOMAZZI),

Гислен ШАРТРОН
(Ghislaine CHARTRON)

Хранилище искусств и ремесел, Лаборатория информационных устройств и средств коммуникации, Франция

Спустя два года после публикации Европейской комиссией рекомендации по открытому доступу к научной информации был преодолен критический порог доступности для пятидесяти процентов статей. Однако этот показатель является средним и выполнение рекомендации Европейской комиссии варьируется от одной страны к другой. Актуальным сейчас является наблюдение различных шагов выполнения и удивление по поводу причин такого неравенства. В целях предложения многих элементов ответа это исследование сравнивает разные уровни выполнения рекомендации в ЕС28.

Вопреки тому, что Европейская комиссия могла бы ожидать в дальнейшем в отношении к коммуникации и рекомендации (затрагивающей открытый доступ и сохранение научной информации в рамках подхода «Горизонт 2020») после ее публикации 17 июля 2012 г., ее выполнение национальными правительствами и научными финансовыми организациями ЕС (Европейский Союз) не привело к стандартизации политики открытого доступа. Эта рекомендация прошла все этапы реализации, касающиеся уровня инициативы, рассматриваемых контентов, периодов эмбарго и т.д.

Прежде всего эта статья предлагает провести сравнение между национальными выполнениями рекомендации Европейской комиссии в ЕС28. Предложенный анализ – хороший пример ее (рекомендации) различных интерпретаций и выполнения. Мы сравниваем планы предпринятых действий и их методы: принудительное введение и национальная рекомендация, делегирование в каждое учреждение и научно-финансовую организацию, национальное консультирование по поводу мнения заинтересованных сторон, отсутствие всякой политики.

МЕТОДОЛОГИЯ

Это исследование проводилось на основе библиографических ресурсов открытого доступа в ЕС28, собранных через поисковую систему BASE, и другой информации из портала OPENAIRE и портала глобального поиска ЮНЕСКО.

ВЫПОЛНЕНИЕ РЕКОМЕНДАЦИИ НА НАЦИОНАЛЬНОМ УРОВНЕ

Несмотря на рекомендации Европейской комиссии, мы обращаем внимание, что есть четыре уровня выполнения: отсутствие национального мандата на открытый доступ и политики, консультирование в процессе применения национальной политики, мандаты финансовых организаций и политика, координированная с помощью рекомендации или закона национальная политика.

Отсутствие национального мандата на открытый доступ и соответствующей политики

Европейскими странами, которые не применяют национальную политику открытого доступа, являются: Румыния, Кипр, Греция, Эстония, Болгария, Мальта, Словакия, Литва, Чехия, Люксембург.

Эти страны представляют некоторые общие характеристики, объясняющие статус-кво в национальном применении европейской политики открытого доступа. Во-первых, все страны (исключая Эстонию, Чехию и Люксембург) имеют внутренние валовые расходы на научно-исследовательские и опытно-конструкторские работы в качестве доли валового внутреннего продукта меньше 1, тогда как самая низкая доля составляет 0, а

* Перевод Lomazzi L., Chartron G. The implementation of the European Commission recommendation on open access to scientific information: Comparison of national policies// ELPUB 2014. Let's put data to use: Digital scholarship for the next generation, 18 th International Conference on Electronic Publishing, 19-20 June, 2014, Thessaloniki, Greece.—http://elpub.scix.net/data/works/att/103_elpub2014.content.pdf

самая высокая – 3,5. Во-вторых, есть страны, публикующие меньше 1 000 научных статей в год, кроме Греции и Чехии. Короче говоря, существует совсем маленькое число заинтересованных лиц на европейской научной сцене.

Можно легко прийти к выводу, что несмотря на осуществимые в дальнейшем бюджетные сбережения благодаря открытому доступу к научным публикациям, эти страны не могут позволить себе создать инфраструктуры и фонды открытого доступа. В некоторых случаях необходимые инфраструктуры существуют, но желание внедрить политику открытого доступа наталкивается на отсутствие осведомленности ученых или недостаточный спрос, вызванный числом опубликованных статей на национальном уровне.

Развитие консультирования относительно применения национальной политики

Четыре европейских страны пока не осуществили скоординированную национальную политику, но находятся на правильном пути. Действительно они начали проводить работу на уровне национального консультирования со всеми заинтересованными сторонами, которая должна привести к предложению законопроекта.

В Польше национальная консультация относительно открытого доступа к публичным ресурсам была инициирована министром администрирования и вопросов оцифровки в 2012 г. Ее цель состояла в том, чтобы определить руководящие принципы политики открытого доступа, которые будут объединены в законопроект, включающий открытый доступ к образовательным, культурным и научным ресурсам и будут финансироваться государством, — «Закон по открытым общественным ресурсам». Опасение не предоставлять открытый «золотой» доступ в долгосрочной перспективе ведет к одобрению «зеленого» открытого доступа.

В Словении закон о НИОКР гласит, что результаты финансируемого государством исследования должны быть доступными. Цель первого периода (2011 — 2014 гг.) *Резолюции по национальной программе НИОКР на 2011-2020 гг.* состояла в том, чтобы начать проводить широкое, на национальном уровне, консультирование с каждой заинтересованной стороной в целях установить некоторые рекомендации для будущего законопроекта, который включал бы и данные. *План национальной программы НИОКР 2011-2020 гг.* также упоминает связь всех национальных архивов в Информационной системе по текущим научным исследованиям (Current Research Information System – CRIS, США)/ Информационной сисстеме по текущим научным исследованиям Словении (Slovenian Current Research Information System – SICRIS).

В Нидерландах, начиная с 2009 г.ректоры университетов ясно высказались в пользу открытого доступа, говоря о средствах поощрения реализации открытого доступа. Нидерландская организация по научным исследованиям (The Netherlands Organization for Scientific Research – NWO), независимая исследовательская организация, которая финансирует исследования и является одним из самых больших голландских финансовых фондов, проводит сильную политику в пользу открытого доступа, в частности в отношении «золотой» дороги путем выделения субсидий на осуществление программы грантов в качестве вознаграждения авторов. В настоящее время нет никакого проекта по реализации политики открытого доступа, а существует только консультирование на национальном уровне.

Во Франции, даже если речь министра высшего образования и научных исследований Женевьевы Фиоразо, произнесенная 24 января 2013 г., и указала на то, что «французское правительство вновь подтверждает свою поддержку принципа открытого доступа к научной информации», выполнение обязательной политики открытого доступа не имеет единодушного одобрения, особенно среди издателей в Секторе социальных и гуманитарных наук ЮНЕСКО (SHS). Национальное консультирование было начато недавно Министерством высшего образования и научных исследований с целью установить, что является оптимальным периодом эмбарго для журналов SHS. В настоящее время, есть пять обязательных политик в отношении депозита (IRSTEA, IFREMER, CIRAD, INRA, INRIA) и две национальные политики с поддержкой финансовых сторон (CNRS, INSERM).

Мандаты финансовых организаций и политика

В настоящее время в Великобритании «золотая» дорога обсуждается больше, чем «зеленая», даже если последняя и не отвергается. Научный совет Великобритании, консорциум семи независимых научных советов, устанавливает политику «золотого» открытого доступа. Эта политика изучалась и сопровождалась промежуточным отчетом, который должен быть пересмотрен осенью 2014 г. Шестнадцать других финансовых организаций также имеют собственную политику в отношении открытого доступа, их список доступен на сайте SHERPA/ROMEO.

В Дании 22 июня 2012 г. пять основных национальных финансовых организаций (Датский совет по независимому исследованию, Датский совет по стратегическому исследованию, Датские национальные исследовательские фонды, Датский фонд по новейшим технологиям и Датский совет по технологии и инновации) решили установить общую политику открытого доступа. Эта политика требует депозитарного хранения цифровой версии научных статей в открытых архивах в течение шести или двенадцати месяцев после принятия статьи. Семь университетов из восемь имеют политику открытого доступа. Однако часто это больше декларация о намерении, чем реальный мандат.

В Финляндии, даже если принцип открытого доступа поощрялся в течение долгого времени, конкретные действия возникли только недавно. В 2011 г. министр образования и культуры запустил проект, названный TTA, с целью создания национальной научной политики по открытому доступу и построения необходимой инфраструктуры. В настоящее время национальный законопроект распространен среди различных заинтересованных сторон, чтобы они смогли его прокомментировать. Этот законопроект рекомендует или «золотую» или «зеленую» дорогу, но не принимает во внимание комбинированные публикации. Было выделено открытое финансирование доступа. Академия наук, которая является главной финансовой стороной, рекомендует ученым публиковаться в журналах открытого доступа как можно чаще.

В Швеции две главных финансовые организации, Шведский научный совет и Шведский научный совет по вопросам окружающей среды, сельскохозяйственным наукам и пространственному планированию (FORMAS), установили мандат на открытый доступ (мандат «зеленого» открытого доступа, касающийся рецензируемых статей для депонирования в открытых архивах в течение шести месяцев после публикации), а Ассоциация выс-

шего образования Швеции (SUHF) рекомендует открытый доступ к 41, входящему в ее состав институту, и поощряет их учреждать собственную политику открытого доступа. Недавно Шведский научный совет (SRC) был назначен министерством в качестве ответственной организации за установление основных руководящих принципов национальной политики в пользу открытого доступа. Первая версия этого сообщения должна была быть опубликована к концу 2014 г.

В Австрии движение за открытый доступ началось в 2009 г. За два года его темп ускорился в связи с получением мандатов ряда финансовых организаций, особенно Австрийского фонда науки (FWF), который рекомендует ученым публиковаться в журналах открытого доступа (гонорары авторам выплачиваются из выделенного фонда) и депонировать электронную версию в открытых архивах в течение двенадцати месяцев после публикации. Австрийская академия наук (OAW) придерживается политики «зеленого» открытого доступа, но также имеет издательство, которое выпускает журналы и книги «золотого» открытого доступа. Другие политики/мандаты учреждений должны быть скоро установлены, но в настоящее время нет никакого выражения потребности относительно осуществления национальной политики по открытому доступу.

Венгрия имеет особенно активную национальную исследовательскую среду, подкрепленную правительством и Венгерским научным фондом исследований (OKTA), который является главной финансовой стороной. Политика OKTA поощряет открытый доступ, требуя, чтобы финансируемые ученые публиковались в журналах открытого доступа и депонировали электронную версию в открытых архивах. Единственное современное правительственное постановление по открытому доступу касается докторской диссертации (№ 33, 7 марта 2007 г.).

Национальная политика, скоординированная в соответствии с рекомендацией

В Бельгии действительно трудно установить национальную политику открытого доступа вследствие федерализма, который явно усложняет координацию между различными региональными научными средами, издавая выпуски, освещающие проблемы заинтересованных сторон и лингвистические проблемы. Однако две главные научные организации— FWO во фламандском сообществе и FNRS во французском сообществе — обе имеют мандат «зеленого» открытого доступа, принятый в 2013 г., который требует депонирования публикаций ученых в открытых архивах. Первым шагом к этому было осуществление национальной политики открытого доступа на основе Брюссельской декларации, которую 22 октября 2012 г. подписали официальные представители министров Валлонского, Брюссельского и Фламандского региона. Эта декларация определяет бельгийскую политику в сфере открытого доступа. Подписавшие ее обязались поощрять открытый доступ к финансируемым государством результатам исследований путем информирования ученых, советуя им сделать их публикации доступными в течение, по крайней мере, шести месяцев (STM) и двенадцати месяцев (SHS) после публикации, изучения возможностей общественных фондов выплатить гонорары за публикацию в открытом доступе, развивая создание и сохранение депозитных инфраструктур, обсуждения рисков и возможностей каждого пути открытого доступа с заинтересованными сторонами. Этот диалог превратился в национальную консультацию, и синдикат издателей собирается подписывать соглашение с университетами, что может привести к периодам эмбарго от шести до двенадцати месяцев и даже больше для публикаций в сфере гуманитарных и общественных науках.

В Ирландии из 7 национальных финансовых организаций четыре имеют мандат относительно открытого доступа (Официальный орган по высшему образованию, Совет по исследованиям в сфере здравоохранения, Ирландский научный совет по науке, инженерному делу и технологиям). 23 октября 2012 г. правительство объявило о том, каковы национальные принципы открытого доступа в *Заявлении о национальных принципах политики открытого доступа*. Среди главных принципов мы нашли обязательство по депозитарному хранению научно-исследовательских, финансируемых государством, публикаций и инициативу публиковаться в журналах открытого доступа. Эта рекомендация благоприятствует «зеленой» дороге, но определенно не устраняет и «золотой» путь. Это соответствует созданию специализированного фонда в целях установления учрежденческих депозитариев и национального портала, в котором не используется никакой определенный фонд для финансирования «золотой» дороги.

В Португалии некоторые инициативы открытого доступа были установлены с 2004 г. Хотя португальское правительство, общественные и частные финансовые организации еще официально не объявили о политике открытого доступа или мандатах, Конференция ректоров португальских университетов (CRUP) рекомендует научным организациям осуществлять мандатную архивную политику для научных публикаций и данных. CRUP доверяет общему правилу единого европейского мандата относительно открытого доступа, который мог привести к отсутствию реализации национального мандата.

В Хорватии есть научное сообщество по открытому доступу, активность которого особенно заметна на примере четырех учрежденческих архивов и одного национального портала, который делает доступными более 250 научных хорватских журналов (HRCAK). В настоящее время в Хорватии нет ни одного мандата финансовой организации в отношении открытого доступа. Документ *Научная и технологическая политика Республики Хорватия в 2007-2010 гг.*, выпущенный Министерством науки, образования и спорта, упоминает, что финансируемые государством результаты исследований должны быть доступными для широкой публики благодаря публикациям или базам данных открытого доступа. 24 октября 2012 г. национальная декларация была доведена до общественности.

Национальная политика, скоординированная в соответствии с законом

В Латвии принятие кабинетом министров национальной программы реформы по выполнению европейской стратегии «Горизонт 2020» не привело к принятию политики открытого доступа или мандатов, выданных финансовыми организациями или правительством на долгий срок. Однако эта программа упоминает обязательство депонировать финансируемые государством научные публикации в архивах (периоды эмбарго от шести месяцев в STM (издательство Science/Technical/Medicine Publishers) и до двенадцати в SHS) и создание субсидий для журналов «золотого» доступа.

Испания была первым государством, с 2011 г. узаконившим открытый доступ с помощью «Ley de la Ciencia, la Technologia y la Innovacación» (Закон о науке, технологии и инновации). Выполнение этого закона не очень пагубно для издателей, поскольку он поддерживает редакционное эмбарго, как упоминается в параграфе 3 статьи 37 данного закона.

В Германии датированный июлем 2013 г. закон о «повисших в воздухе» и недоступных работах, включает пункт об открытом доступе. Пункт дает авторам право вторичной публикации. Это позволяет брать аналогичную, но некоммерческую публикацию автора в течение двенадцати месяцев после принятия статьи в STM и SHS. Это право применяется, если научная работа является финансируемой государством и если статья принята в журнале, т. е. опубликована, по крайней мере, два раза в год. Это решение подтверждается своим превосходством в контракте.

В Италии в марте 2013 г. президенты основных научных организаций, связанные с Конференцией ректоров итальянских университетов, подписали декларацию в пользу открытого доступа. В октябре 2013 г. представитель законодательства выступил с замечанием относительно открытого доступа, касающимся постановления закона о сохранности и восстановления культурных ценностей. Однако, тогда как начальный законопроект планировал открытый доступ к статьям в течение шести месяцев после публикации, законопроект, принятый 8 октября 2013 г., требует периодов эмбарго в 18 месяцев в STM и 24 месяца в SHS, а книг вообще не касается. Эта модификация первой версии законопроекта является следствием важной работы по лоббированию, которое было проведено частными итальянскими издателями, полагающими, что шестимесячные периоды эмбарго недостаточны для обеспечения (гарантии) экономической жизнеспособности публикаций.

В качестве заключения важно знать, что это исследование является срезом ситуации в определенное время. Действительно, данные развиваются со временем и должны постоянно актуализироваться.

Однако в конце этого исследования мы замечаем, что дисбаланс появляется с самого начала выполнения рекомендации Европейской комиссии. Это приводит нас к вопросу, кто на самом деле точно выигрывает от открытого доступа, страны, которые лидируют в мире, внося научный результат, или те страны, что остаются позади? Чтобы ответить на этот вопрос, необходимо рассмотреть две особенности: определенный язык производства статей и научное закрепление дисциплины или в гуманитарных науках, или в естественных науках. Как следствие, эта проблема будет темой дальнейшего исследования относительно будущего национальной публикации (не на английском языке) в контексте рекомендации ЕС.

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 10.03.2015 г.
Бумага офсетная. Формат 60x841/8. Гарн. литер. Печать цифровая
Усл. печ. л 4,50 Уч.-изд. л. 5,11 Тираж 38 экз.

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