ОТКРЫТЫЙ ДОСТУП
И ОТКРЫТЫЕ АРХИВЫ ИНФОРМАЦИИ

УДК 026.06

В. М. Московкин

Базы данных научной информации и онлайновые поисковые инструменты: использование для управления знаниями

Проанализированы основные зарубежные библиографические базы данных и поисковые инструменты, включая полнотекстовые БД открытого доступа. Представлен опыт управления знаниями на основе онлайновых инструментов.

Ключевые слова: научная информация, открытый доступ, полнотекстовые базы, базы данных, электронные архивы, репозитарии, онлайновые научные журналы, Google Scholar, Scirus, Google Books, Google Patents, Google Analytics, университетский вебометрический бенчмаркинг, вебометрический рейтинг.

В 1960–1990-е гг. в мире существовали две глобальные конкурирующие между собой базы данных научной информации – Института научной информации США (Филадельфия) и Всесоюзного института научной и технической информации (Москва). Они были построены на разных принципах, но имели приблизительно одинаковый годовой входящий документопоток (около 1 млн научных документов).

Деятельность ВИНИТИ способствовала проведению широкомасштабных научных исследований на высочайшем научном уровне, что содействовало научно-техническому прогрессу СССР в период холодной войны. Большую роль играл институт внештатных референтов – известных исследователей в различных узкоспециализированных областях, который обеспечивал налаженную систему реферирования статей, поступающих со всего мира.

После распада СССР, в условиях финансово-экономического кризиса, роль ВИНИТИ начала уменьшаться; реферативные журналы значительно подорожали (следовательно, стали менее доступными), а их качество ухудшилось, так как разрушилась система реферирования. Постепенно реферативные журналы перестали быть атрибутом научных исследований (например, современное поколение экономистов даже не знает о существовании журналов «Экономика промышленности» и «Организация управления»). Этому во многом содействовало развитие Интернета в конце 1990-х – начале 2000-х гг.

Исследователи, научные менеджеры, информационные и библиотечные работники стали искать новые возможности для поддержки исследовательской и информационно-библиотечной деятельности. Началось создание и использование онлайновых баз данных научной информации. Но если для зарубежных организаций обращение к таким БД, а крупнейшие из них – Web of Science (базы данных ISI– *Institute of Science Information*, США), SCOPUS, – это неотъемлемая составляющая работы, то для наших научных учреждений подобные БД слишком дороги: годовая подписка 25–35 тыс. долларов. Следует также иметь в виду, что эти БД охватывают только журнальные коллекции (базы данных ISI– «SCI», «SSCI» и «A&HCI» – отражают около 9 тыс. журналов мира с самым высоким рейтингом, SCOPUS – около 30 тыс.).

В начале XXI в. возникла мощная поисковая некоммерческая система научной информации, которая ведет поиск по всему Интернету. Это Google Scholar[**1**, **2**]. Она охватывает те же исходные ресурсы, которые есть в базах данных ISI и SCOPUS, а также менее значимые научные ресурсы, патенты, материалы, размещенные в электронных архивах открытого доступа и онлайновых научных журналах, и др. Достоинства этой поисковой машины – разработанные Google специальные алгоритмы расчёта цитируемости документов (опция «by cited»), а также возможности поиска научных документов на сайтах научных организаций (в 2004 г. испанская киберметрическая лаборатория использовала Google Scholar при разработке вебометрического рейтинга университетов и научно-исследовательских центров мира).

Google Scholar позволяет вести поиск с учетом различных логических операторов, включая расширенный поиск с точной фразой, ограничением по временному интервалу и области знаний (используются семь широких областей знаний) и др.

Опыт показывает, что Google Scholar очень удобен при бенчмаркинге университетской публикационной активности, расчете укрупненных публикационных структур (распределении публикаций по областям знаний) и их типизации, идентификации научных фронтов и кластеров научных публикаций [**3**]. Но главное его преимущество состоит в том, что он решает проблему составления литературного аналитического обзора при

проведении любых исследовательских и диссертационных работ (поиск научных документов по ключевым словам).

В отличие от другой мощной поисковой системы научной информации – Scirus, Google Scholar включает много русскоязычных научных документов. Сюда, по крайней мере, входят все статьи, размещенные в русскоязычных репозитариях и научных журналах открытого доступа.

К конкурентам Google Scholar можно отнести только упомянутый выше инструмент Scirus. Несмотря на то, что разработчики называют его всеохватывающим (*comprehensive*), он не ведет поиск по всему Интернету, как это делает Google Scholar, а имеет определенный, хотя и очень обширный перечень научных информационных ресурсов: около 140 млн документов с сайтов с доменами *edu* (университетские домены), 40 млн – *org*, 39 млн – с доменами *gov*, 38 млн – *com*, 23 млн – *ac.uk* (британские академические организации – университеты), свыше 136 млн других релевантных STM и университетских сайтов со всего мира, около 40 мощных научных порталов и платформ (например 627 тыс. препринтов в Arxiv.org), 3,6 млн документов из цифровых архивов (около 200 репозитариев). Этот поисковый инструмент охватывает около 410 млн научных документов – научные статьи, книги, диссертации, курсы лекций, препринты, патенты, домашние страницы ученых, веб-информация.

Scirus запущен при поддержке издательства «Elsevier» (неудивительно, что больше всего откликов – с платформы Science Direct, которая является онлайновой платформой этого издательства).

Scirus, как и Google Scholar, в процессе поиска научных статей отфильтровывает ненаучные статьи и ищет исключительно рецензируемые. Обе поисковые машины идут глубже, чем первые два уровня веб-сайтов, поэтому обнаруживают гораздо больше релевантной информации.

В отличие от Google Scholar, в Scirus предусмотрены более дробная градация поиска ключевых слов в различных местах метаданных и всего документа (название статьи, название журнала, ключевые слова статьи, ISSN, имя автора, весь документ), поиск по виду документа (книги, статьи, препринты, патенты, рефераты, тезисы и диссертации, домашние страницы авторов или организаций, обзоры, конференции), по форматам файлов (pdf, ppt, tex, HTML, ps, word).

При расширенном поиске автоматически выдается информация о количестве журнальных источников, предпочтительных веб-ресурсах, других веб-ресурсах, а также типах файлов. От этих количественных распределений через гиперссылки сразу же можно просматривать найденные научные документы. Так же, как и в Google Scholar, реализованы логические операторы и возможность поиска с точной фразой. В отличие от семи широких научных областей в Google Scholar, в Scirus используются девять широких научных областей, но сравнение классификаций не в пользу Scirus:

- классификация научных областей в Scirus – Agriculture and Biological Sciences; Astronomy; Chemistry and Chemical Engineering; Computer Sciences; Earth and Planetary Sciences; Economy, Business and Management; Environmental Sciences; Engineering, Energy and Technology; Languages and Linguistics;
- классификация научных областей в Google Scholar – Biology, Life Sciences and Environmental Sciences; Business, Administration, Finance and Economics; Chemistry and Material Sciences; Engineering, Computer Sciences and Mathematics; Medicine, Pharmacology and Veterinary Sciences; Physics, Astronomy and Planetary Sciences; Social Sciences, Art and Humanities.

В первой классификации отсутствуют физика, математика, большинство социальных наук. Вторая классификация более логична, несмотря на меньшее количество градаций областей знаний.

Помимо Google Scholar исследователям и информационно-библиотечным работникам следует обратить внимание на поисковые системы Google Books и Google Patents. Первая охватывает свыше 10 тыс. издателей и авторов, публикующих книги на более чем 35 языках (свыше 5 млн книг); позволяет вести поиск по языкам, названиям, авторам, издателям, предметным областям, ISBN и интервалам времени. Особый интерес представляет возможность идентификации и систематизации старопечатных изданий XV–XVI вв. благодаря широко объявленной и реализуемой компанией Google амбициозной программе по всеохватывающей оцифровке книг, особенно тех, на которые из-за давности лет не распространяется авторское право (компания Google заключила договоры с более чем 40 крупнейшими публичными и университетскими библиотеками США).

Google Patents охватывает свыше 7 млн патентов на изобретения и торговые марки, входящие в базу данных патентного ведомства США. У этой системы широкие возможности для поиска (даты, патентные классификации, авторы, организации и т.д.).

Существуют узкоспециализированные коммерческие онлайновые БД научной информации. Десять крупнейших издательств научной периодики формируют базы данных ISI (9 тыс. журналов), имеют собственные онлайновые

платформы со своими поисковыми интерфейсами: «Elsevier» – платформа Science Direct, «Springer» – платформа Springer Link, «Wiley» – платформа Wiley InterSciences, «SAGE Publication» – платформа SAGE Journals Online и т.д. На всех этих платформах заказ полнотекстовых документов (статей) будет стоить в среднем от 30 до 40 долларов за одну статью.

Наиболее известная и крупная – платформа ScienceDirect, охватывающая более 2,5 тыс. журналов и около 9,5 млн полнотекстовых статей; ежегодный прирост – около 500 тыс. статей. Старейшие журналы оцифрованы с начала XIX в. (например журнал «Lancet», 1823 г.). С 2003 г. можно размещать дополнительные материалы (видео, аудио). Для гостей платформы предусмотрен Quick Search, разрешающий бесплатный просмотр библиографических описаний и аннотаций статей. Кроме того, можно просматривать эти же документы для библиографических списков. Для лицензированных пользователей предусмотрен Basic and Advance Search, который позволяет не только получать полнотекстовые документы, но и использовать аналитические инструменты SCOPUS: импакт-факторы журналов, индексы цитируемости статей, построение сетей (связей) цитирования и др.

Следует иметь в виду и основные международные отраслевые БД научной информации: PubMedCentral, Public Library of Science, Medline, Inspec, Econlit, EBSCO Hostweb и др., а также международную БД диссертаций и их тезисов (ProQuest Dissertations and Theses). Отмечу, что отраслевые БД могут быть как коммерческие, так и открытого доступа.

Полнотекстовые базы данных открытого доступа

Возникшее в конце 1990-х–начале 2000-х гг. международное движение за открытый доступ (ОД) к научному и гуманитарному знанию, его многочисленные инициативы и декларации привели к созданию глобальных сетей электронных архивов ОД (репозитариев) и онлайновых научных журналов [4]. Для каталогизации первой сети и поиска в ней были созданы два регистратора – ROAR (Registry of Open Access Repositories, Саутгемптонский университет) и DOAR (Directory of Open Access Repositories, Нотингемский университет), для второй – регистратор DOAJ (Directory of Open Access Journals, университет г. Лунд, Швеция). Динамика наполнения этих регистров ресурсами ОД за последние 5 лет отражена в табл. 1. С каждым годом динамика будет только нарастать, так как сейчас из всего «научного выхода», поддерживаемого 25 тыс. научных журналов, только 15 % находится в полнотекстовом открытом доступе.

Из первых двух регистров предпочтение следует отдать ROAR. В нем значительно лучше разработан интерфейс. По объемам наполнения эти регистры приблизительно одинаковы, так как организации обычно регистрируют свои репозитарии и в том, и в другом. В то же время в DOAR реализован ряд аналитических инструментов, отсутствующих в ROAR (они необходимы для построения обобщенных распределений в виде диаграмм и таблиц по различным характеристикам: страны, виды документов, типы репозитариев и т. д.)

Таблица 1

Динамика наполнения регистров ресурсами открытого доступа

ОД-регистры	**Дата**	**Количество ресурсов**	**Прирост за период с 28 (29).10.2009 г. по 28.10.2011 г., в %**
ROAR	11.05.2006 25.01.2008 28.10.2009 08.11.2010 28.10.2011	658 989 1 511 2 049 2 486	64,5
DOAR	26.05.2009 29.10.2009 08.11.2010 28.10.2011	1 394 1 509 1 815 2 128	41,0
	15.01.2008	3 095/1707244	

DOAJ	29.10.2009 08.11.2010 28.10.2011	392/3210175 842/4842687 250/655382	65,1/104,2

В ROAR предусмотрен поиск репозитариев по странам (табл. 2), содержанию, виду программного обеспечения. Наиболее популярное открытое программное обеспечение – DSpace (960 репозитариев на 28.10.2011 г., 771 – на 18.11.2008 г.) и EPrints (416 репозитариев на 28.10.2011 г., 366 – на 18.11.2008 г.), типу репозитариев (журнальные публикации, диссертации, демонстрационные, обучающие и др.). Кроме того, предусмотрена сортировка репозитариев ОД по активности их наполнения. На сайте ROAR организации, имеющие собственные репозитарии, в разделе ROARMAP могут регистрировать свою институциональную политику открытого доступа.

Анализ данных табл. 2 показывает значительное увеличение количества репозитариев в Бразилии (39,7 %). Среди развитых стран максимальный прирост наблюдался в Великобритании, а минимальный – в Австралии и Японии. США в этом регистре стартовали с очень большим количеством репозитариев и поэтому прирост был относительно небольшим (11,8 %). Пустые клетки в таблице означают, что в 2008 г. мы опирались не на TOP–14, а на TOP–7.

Таблица 2

Распределение репозиториев в ROAR по странам

Страна	**Количество репозитариев ОД**		**Прирост, %**
	08.11.2008	**28.10.2011**	
США	350	397 (401)	11,8
Великобритания	185	218 (201)	17,8
Япония	131	138 (135)	5,3
Германия		134 (148)	
Бразилия	78	109 (60)	39,7
Испания		95 (77)	
Индия		76 (53)	
Тайвань		72 (58)	
Канада	62	70 (58)	12,9
Италия	62	69 (65)	11,3
Польша		67 (75)	
Австралия	63	66 (64)	4,8
Швеция		65 (46)	
Франция		60 (64)	
Россия	33	35 (13)	6,1
Украина	17	35 (25)	105,9
Беларусь	0	1 (1)	0

Репозитарии располагаются в ROAR в порядке убывания их записей, поэтому в начале – крупные порталы, цифровые библиотеки и даже поисковые машины научной информации, имеющие более 1 млн записей: PubMedCentrale, Networked Digital Library of Theses and Dissertations Union Catalog, Humanities Text Initiative, Citeseerx Scientific Literature Digital Library and Search Engine (CiteSeerX), RePEc (Research Papers in Economics) и др. В этот же список входит и Google Scholar – универсальный поисковик научной информации в Интернете. Он зарегистрирован в ROAR 4 мая 2006 г., но общее количество записей для него определить невозможно.

Из ROAR можно сразу же выходить на сайты конкретных репозитариев OD. В нём реализован важный аналитический инструмент, позволяющий просматривать динамику наполнения репозитария, предусмотрена также функция графического анализа данных.

В DOAR возможен поиск репозитариев по областям знаний и языкам (эти функции отсутствуют в ROAR), а также и по другим параметрам, присутствующим и в ROAR. В DOAR хорошо и наглядно реализованы инструменты статистического и графического анализа по всей БД этого регистра (*statistical charts*).

В отличие от ROAR размер (количество записей) репозитариев в DOAR обновляется не оперативно, отсутствуют аналитические инструменты, позволяющие просматривать и изучать динамику и интенсивность пополнения конкретного репозитария.

В DOAJ предусмотрен поиск журналов по их названиям и областям знаний. Например, при выборе первой области знаний «Agriculture and Food Sciences» возникает пять дополнительных узких областей с указанием количества журналов в них. По гиперссылкам можно просматривать профили конкретных журналов: ISSN, EISSN, предметная область, издатель, страна, язык, ключевые слова, год первого выпуска журнала.

На сайте DOAJ реализованы функции просмотра новых поступлений журналов, регистрации новых журналов; дана полезная информация для пользователей, библиотек, издательств и авторов.

В 2009 г. DOAJ получил от SPARC Europe четвертую премию за выдающиеся достижения в области научных коммуникаций. На сайте DOAJ в системе GEOVISITE реализована функция, показывающая одномоментное количество пользователей, зашедших на этот сайт, по всем странам (в среднем 3–4 тыс.).

Уникальность этого регистра привлекла к нему внимание крупнейших издательств (Springer, Sage, BioMed Central), Европейской организации по ядерным исследованиям (CERN), цифровой библиотеки общества Макса Планка (Max Plank Digital Library), Британского финансирующего агентства (UK Science and Technology Facilities Council), которые в период с марта 2009 г. по февраль 2011 г. в рамках FP7 выполняли проект «Study of Open Access Publishing» (SOAP, координирующая организация – CERN).

В рамках этого проекта из 4 032 журналов, входящих в DOAJ (июль 2009 г.), были отобраны все англоязычные журналы (2 838) и их издатели (1 809), чтобы проанализировать все возможные количественные распределения. В базу данных Scopus (2009 г.) входило 38 % этих журналов, в ISI-JCR (2008 г.) – 8 %.

В этой выборке были идентифицированы 14 крупнейших издателей, выпускающих 616 журналов (36 096 статей в год), и получена информация о их политике в области авторского права и доходах.

Рассмотренные три регистра научно-информационных ресурсов полностью охватывают все полнотекстовые базы данных открытого доступа.

Управление исследованиями, электронными архивами и библиотечными коллекциями на основе онлайновых инструментов

Для оценки результативности научных исследований, планирования индивидуальных публикационных стратегий, управления электронными архивами и библиотечными коллекциями удобно использовать различные онлайновые инструменты. Первый из них – запущенная при поддержке компании Scopus онлайновая платформа SCIMAGO. Она позволяет вести поиск журналов, входящих в систему Scopus, определять их текущие импакт-факторы и другие наукометрические показатели. Эта платформа уже в течение нескольких лет используется в Белгородском государственном национальном исследовательскои университете (НИУ «БелГУ») для рейтинговой оценки научной активности докторов и кандидатов наук.

Публикация в журнале, входящем в систему Scopus, оценивается в баллах по формуле – 500 IF/IFmax, где IF – текущий импакт-фактор журнала для данной журнальной категории, IFmax – максимальный импакт-фактор среди журналов этой категории. Таким образом, если ученый опубликовал статью в журнале, входящем в среднеимпактную зону для определенной категории, то он получает за нее 200–300 баллов. В то же время публикация в отечественных журналах, включённых в список ВАК, оценивается всего 15 баллами. Внедрение этой формулы в практику университетского научного менеджмента в начале 2010 г. дало ученым НИУ «БелГУ» стимул для подготовки статей на английском языке.

С целью мониторинга и управления англоязычными публикациями ученых НИУ «БелГУ» в университетском электронном архиве ОД создана коллекция таких публикаций. Для ее формирования мы предложили эффективную процедуру, основанную на поисковых возможностях Google Scholar [**3**]. Это избавило библиотечных работников от необходимости просматривать огромный объем списков трудов ученых НИУ «БелГУ» и связываться с ними для уточнения недостающей информации.

Наши эксперименты по построению университетских публикационных структур показали, что публикации университетских ученых хорошо идентифицируются, когда в Google Scholar ведется расширенный поиск с точной фразой по названиям университетов. Если метаданные публикаций хорошо структурированы, то их поиск очень эффективен. В этом случае в первую очередь мы получаем ссылки на статьи из журналов, размещенных на онлайновых платформах крупнейших издателей: Elsevier, Springer, Wiley, Emerald и др. Это как раз те журналы, которые входят в базы данных Web of Science и SCOPUS. Благодаря протоколу по сбору метаданных инициативы «Открытые архивы» (PMHOAI), Google Scholar хорошо индексирует статьи, размещенные в репозитариях ОД. Он также практически полностью охватывает журналы открытого доступа, входящие в DOAJ (за исключением тех, где плохо структурированы метаданные).

Бенчмаркинг публикационной активности ученых НИУ «БелГУ» показал, что за последние пять лет количество

публикаций на английском языке увеличилось на порядок. Этому способствовали мощная поддержка естественно-научных фундаментальных исследований (статус инновационного, а позднее исследовательского университета), создание наноцентра коллективного пользования, стимулирующие баллы за англоязычные публикации, принятие Белгородской декларации об открытом доступе к научному знанию и культурному наследию и создание DSpace-репозитария.

В настоящее время Google Scholar определяет 1 260 англоязычных публикаций ученых НИУ «БелГУ» (тестирование названия университета «Belgorod State University» в строке «With the exact phrase»). В то же время использование оператора site:bsu.edu.ru в Google Scholar определяет 1 280 публикаций на сайте НИУ «БелГУ» (10 ноября 2011 г.), из которых большая часть – русскоязычные. По сравнению с 2006 г. эти показатели выросли на порядок.

Запуск англоязычной коллекции значительно улучшит показатель, который используется в качестве четвертого индикатора (SCHOLAR) при расчете глобального вебометрического рейтинга университетов. В этом случае следует ожидать значительного повышения мирового рейтинга НИУ «БелГУ». Предыдущий резкий скачок рейтинга произошел в июле 2009 г. после запуска DSpace репозитария ОД БелГУ.

Рост числа англоязычных публикаций, представленных pdf-файлами, также улучшит и третий индикатор глобального вебометрического рейтинга НИУ «БелГУ» (RICHFILES), который учитывает количество файлов различных форматов (doc, ps, ppt, pdf).

Мы отслеживаем все показатели в рамках регулярного университетского вебометрического бенчмаркинга на примере университетов Приграничного белорусско-российско-украинского университетского консорциума [**5**].

Таким образом, онлайновая платформа SCIMAGO, поисковая машина Google Scholar и вебометрический рейтинг университетов (www.webometrics.info) – хорошие инструменты университетского бенчмаркинга и менеджмента знаний. К ним следует отнести и четвертый онлайновый инструмент – Google Analytics, который позволяет отслеживать посещаемость сайтов и востребованность находящейся на них информации.

При создании коллекций англоязычных публикаций следует придерживаться издательской политики по самоархивированию и авторскому праву, размещенной на сайте проекта ROMEO [**6**]. Многие издательства разрешают выставлять в репозитариях ОД, на институциональных и персональных сайтах только авторские pdf-файлы статей; кроме того, следует учитывать и издательские эмбарго на размещение статей в открытом доступе. В то же время практика работы с Google Scholar показывает, что нередко эти правила игнорируются. Например, мы часто видим метаданные статьи на онлайновой платформе издателя, при этом доступ к ее полному тексту требует подписки или оплаты (в среднем 30–40 долларов), но тут же справа может быть ярлык издательского pdf-файла статьи – это говорит о том, что автор статьи разместил ее на институциональном или персональном сайте.

Эффективность поисковой машины Google Scholar подтвердил эксперимент, проведенный в Стокгольмском университете [**7**]. Тридцати двум выпускникам, специализирующимся в области социальных наук, было предложено в течение двадцати минут провести поиск релевантной (отобранной студентами) научной литературы для выпускных работ на основе Metalib (поиск по более чем 200 библиографическим БД, доступным в Стокгольмском университете) и Google Scholar. В результате с помощью Google Scholar было найдено в два раза больше релевантной научной литературы, чем с помощью Metalib.

Опыт работы НИУ «БелГУ» (семинары, тренинги, консультации) подтверждает эти результаты, что ставит под сомнение закупку университетами дорогих библиографических БД. На наш взгляд, целесообразнее закупить одну ключевую базу данных (например SCOPUS для исследовательского университета), а высвободившиеся средства направить на обучение студентов и преподавателей работе с Google Scholar.

Списоки сточников

1. **Jacso P.,** Google Scholar: The Pros and Cons // Online Information review. – 2005. № 29, 2. – PP. 208–214.

2. **Noruzi A.,**Google Scholar: TheNewGenerationofCitationIndexes // Libri. – 2005. – № 55. – PP. 170–180.

3. **Московкин В. М.** Возможности использования поисковой машины Google Scholar для оценки публикационной активности университетов / В. М. Московкин // НТИ. Сер. 1. Организация и методика информационной работы. – 2009. – № 7. – С. 12–16.

4. **Московкин В. М.** Международное движение по открытому доступу к научному и гуманитарному знанию:

опыт для постсоветских стран / В. М. Московкин, Л. В. Верзунова // Информационные ресурсы России. – Москва, 2007. – № 1. – С. 14–18.

5. **Московкин В. М.** Вебометрическая оценка публикационной активности университетов: влияние Белгородской декларации /В. М. Московкин // НТИ. Сер. 1. Организация и методика информационной работы. – 2010. – № 2. – С. 12–16.

6. **Московкин В. М.**Институциональные политики открытого доступа к результатам научных исследований / В. М. Московкин // Там же. – 2008. – № 12. – С. 7–11.

7. **Haya G., Nygren E., and Widmark W.** Metalib and Google Scholar: A User Study // Online Information Review. – 2007. – № 31, 3. – PP. 365–375.