О присутствии академических библиотек в научном веб-пространстве

On the place of academic libraries in the web-space of science

А. Е. Гуськов, Д. В. Косяков, Е. С. Быховцев
Государственная публичная научно-техническая библиотека СО РАН,
Новосибирск, Россия

Andrey Guskov, Denis Kosyakov and Egor Bykhovtsev
State Public Library for Science and Technology of the Russian Academy of Sciences Siberian Branch,
Novosibirsk, Russia

В этой статье проводится анализ вебометрических показателей сайтов академических библиотек, их сравнение друг с другом, а также с другими институтами ФАНО. Дополнительно проведено сравнение наиболее крупных распространителей научного контента в России посредством открытого доступа: КиберЛенинка, Mathnet, РГБ, РНБ и другие.

Web-metric indicators for academic libraries' www-sites are analyzed and compared to each other and other FASO RUSSIA institutions. Besides, significant open access scientific content distributors are compared, namely CyberLeninka, Mathnet, Russian State Lirary, National Library of Russia, etc.

Введение

В библиотечной среде уже второй десяток лет идут активные разговоры о переводе ресурсов в электронный формат. Но как далеко удалось в этом продвинуться и насколько быстро мы идем по этому пути? Авторы этой статьи попытались приблизиться к ответу на этот вопрос посредством вебометрического исследования сайтов академических библиотек. В основе этих исследований лежит измерение и сопоставление технических показателей сайтов, таких как количество страниц, документов или внешних ссылок.

Популярность это направление вебометрики приобрело благодаря работам группы Cybermetrics Lab под руководством Исидро Агийо, результатом которых стал постоянно обновляемый вебометрический рейтинг университетов, исследовательских центров, клиник, бизнес-школ и открытых архивов (проект Webometrics Ranking of World Universities: www.webometrics.info). На наш взгляд, рейтинг достаточно адекватно отражает эффективность научной деятельности организации, что подтверждается сравнением с авторитетными рейтингами высших учебных заведений. Важно отметить, что, хотя этот рейтинг и вызывает определенные нарекания, у него есть одно неоспоримое достоинство – он стимулирует научные организации публиковать материалы в открытом доступе.

Методика сбора данных

Большинство вебометрических исследований опираются на общий набор показателей сайта, основывающихся на использовании коммерческих поисковых систем для измерения параметров. Мы также рассматриваем эти характерные индикаторы:

- количество страниц сайтов, проиндексированных поисковой системой (размер сайта) – индикатор степени «присутствия» организации в Интернете;
- количество полнотекстовых документов (рассматриваются форматы файлов pdf, doc(x), ppt(x), реже – ps) – индикатор «открытости» организации, объёмов результатов научных исследований в виде академических статей, препринтов, отчётов, учебных материалов, размещённых в общедоступном пространстве;
- количество «академических» материалов, проиндексированных специализированной поисковой системой Google Scholar, – индикатор качества выставленных в общий доступ материалов;
- количество ссылок с других сайтов на страницы исследуемого сайта (входящие ссылки) – индикатор степени признания уровня организации и её научных результатов в обществе; также рассматривается альтернативный показатель – тематический индекс цитирования Яндекс.

Для регулярного сбора этих показателей с большого числа веб-сайтов была создана ежемесячно пополняемая общедоступная база данных вебометрических индикаторов академических сайтов, размещённая по адресу http://webometrix.ru, которая может стать инструментом для планирования мероприятий по улучшению представления научной организации в Интернете и основой для составления вебометрических рейтингов сайтов. С ее помощью были получены ежемесячные значения вебометрических индикаторов с января 2015 г. по март 2016 г. по сайтам всех научных организаций, подведомственных ФАНО, в том числе 8 научных библиотек: БЕН РАН, БАН, ЦНБ УрО РАН, ГПНТБ СО РАН, ЦНБ ДВО РАН, ФБ РАМН, ЦНСХБ, СибНСХБ.

Характеристика сайтов библиотек

Среди рассматриваемых библиотек у четырех (ГПНТБ СО РАН, ЦНСХБ, ЦНБ ДВО РАН, БЕН РАН) сайты можно отнести к крупным (суммарное количество страниц на сайтах соответствующих доменов более 200 тыс. страниц. ЦНБ УрО РАН и БАН обладают средними по размерам веб-ресурсами (20-40 тыс. страниц). Общие характеристики вебометрических параметров сайтов библиотек представлены на рис. 1. По осям указаны доли организации в суммарных значениях всех библиотек.

Динамика развития сайтов

Размеры веб-сайтов

Одной из вебометрических характеристик сайта является его размер, который чаще всего измеряется как количество страниц, проиндексированных поисковыми системами. Любопытно, что с точки зрения основных (для России) поисковиков Google и Яндекс ситуация качественно различается.

Данные индекса Google (рис. 2) говорят о стагнации, т.к. размеры сайтов, в целом, остаются неизменными. При этом наблюдаются отдельные всплески, которые можно объяснить спецификой работы поискового робота. В определенный момент он находит на сайте новые для себя страницы (в случае с ГПНТБ СО РАН это были, предположительно, электронные каталоги Ирбис) и заносит их в свой индекс, а через некоторое время они проходят через другой этап обработки и по каким-то причинам исключаются из индекса.

Данные Яндекса (рис. 3), наоборот, говорят о том, что размеры почти всех веб-сайтов библиотек имеют положительную динамику. Причем некоторые веб-сайты за год с небольшим увеличили свой размер на порядок, что представляется нам маловероятным. Рассматривая случай ГПНТБ СО РАН, едва ли можно предположить, что за последний год количество страниц на сайте увеличилось с 20 до 200 тысяч. По всей видимости, в этом случае наблюдаются последствия изменения алгоритмов индексирования Яндекс, благодаря чему в поисковый массив стало попадать больше ресурсов. Такая же картина наблюдается и для других сайтов научных организаций.

Кроме того, различия в алгоритмах индексации видны еще и при сравнении рангов сайтов. Так на графиках лидеры (ЦНСХБ и ГПНТБ СО РАН) меняются местами, а ЦНБ ДВО РАН с лидирующих позиций в Google переносится на предпоследнее место в Яндекс.

Анализируя всю совокупность данных, мы пришли к выводу, что реальная ситуация ближе к той, которую показывает Google. Рассматривая данные с октября 2015 г. по март 2016 г. (где отсутствуют «выбросы»), мы заметили, что общий объем размеров сайтов организаций ФАНО медленно растет (доля НИИ увеличилась с 60,7% до 63,2%), а сайтов библиотек –уменьшается с 16% до 15%. Таким образом, при ожидаемом росте объема академического веб-пространства, доля академических библиотек в нем постепенно снижается.

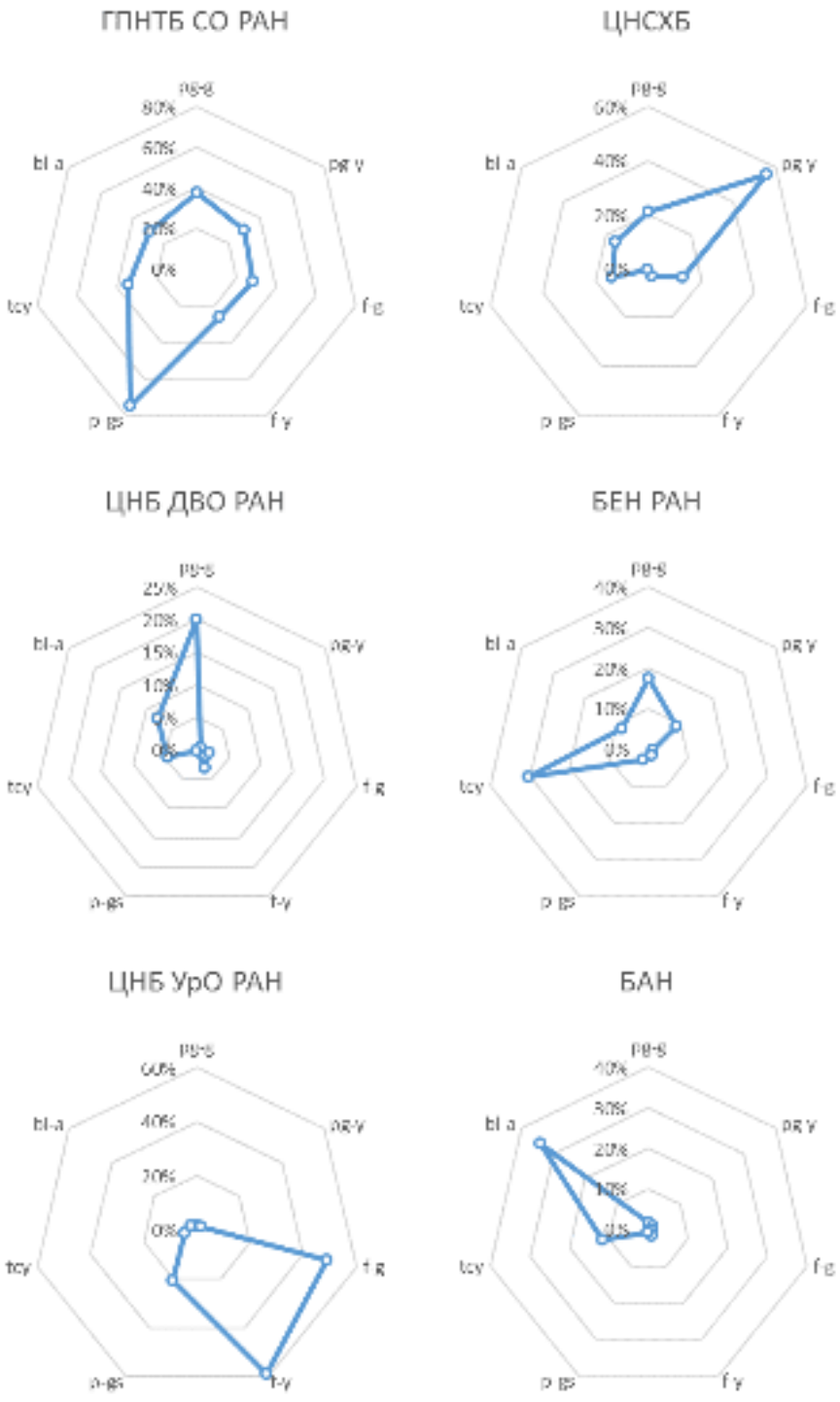

Рисунок 1. Характеристики сайтов шести ведущих библиотек.
По осям полярной диаграммы доля показателя библиотеки в суммарном значении показателя по всем библиотекам. Показатели: pg-g – кол-во страниц сайта по данным Google, pg-y – кол-во страниц сайта по данным Яндекс, f-g – кол-во файлов (полнотекстовых документов) по данным Google, f-y – кол-во файлов по данным Яндекс, p-gs – кол-во документов в индексе Google Scholar, tcy – тематический индекс цитирования Яндекса.

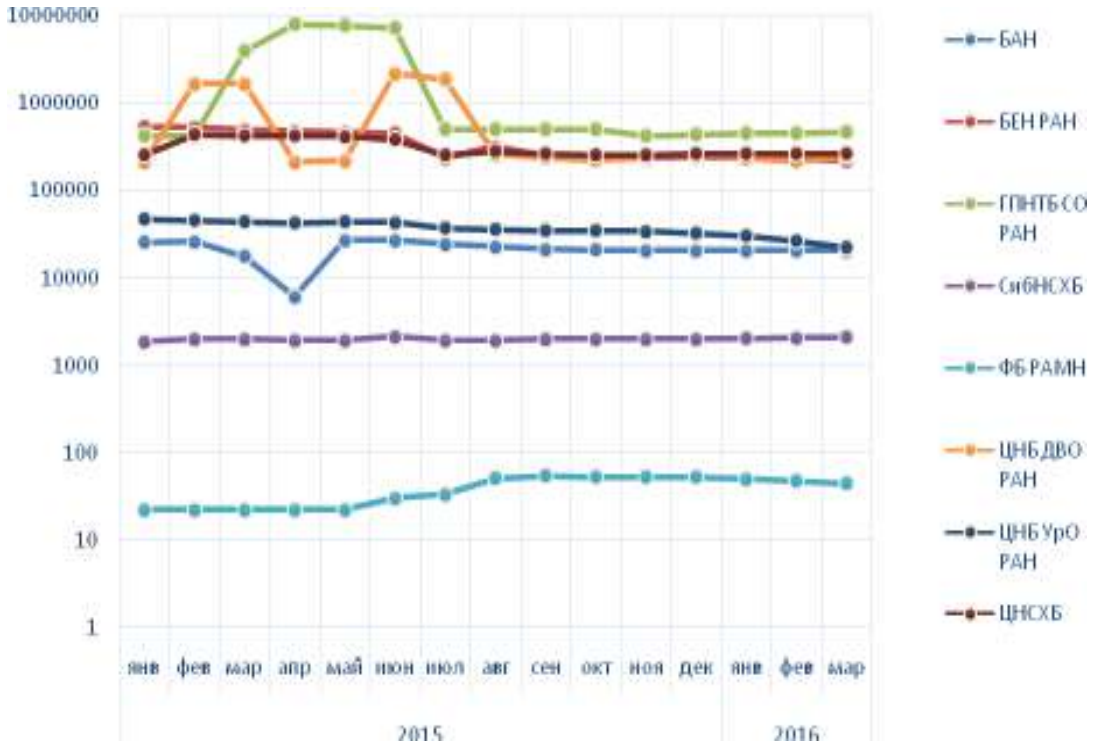

Рисунок 2. Динамика размеров веб-сайтов по данным Google
(кол-во страниц, логарифмическая шкала)

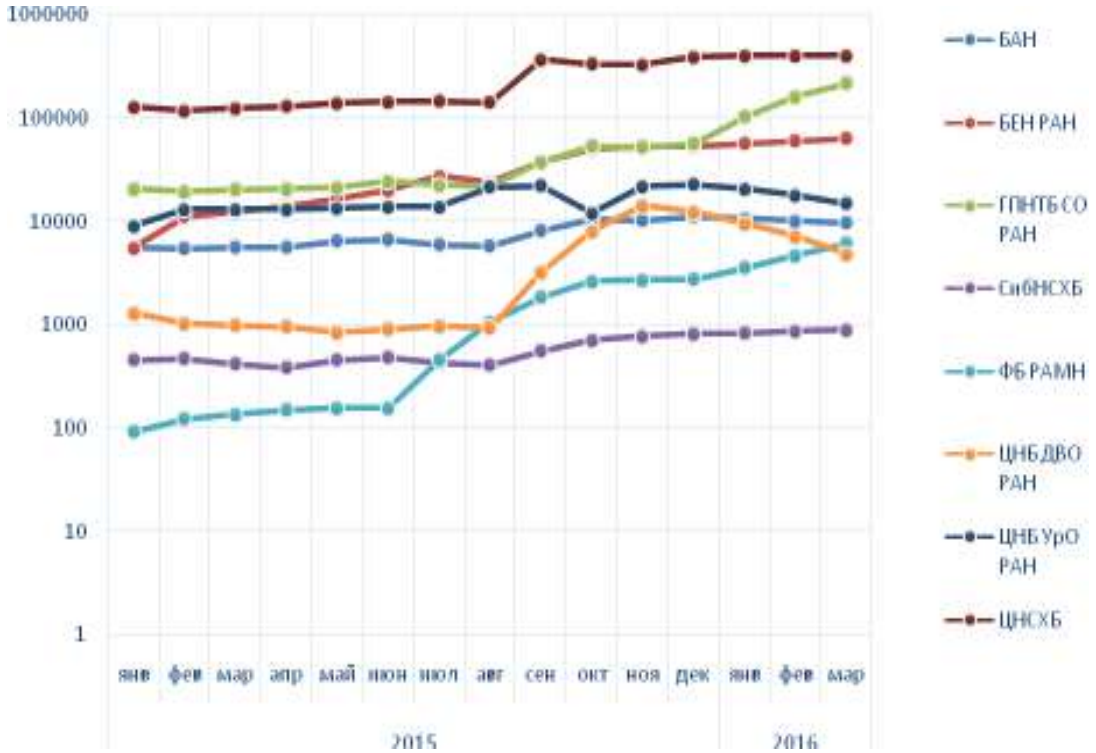

Рисунок 3. Динамика размеров веб-сайтов по данным Яндекс
(кол-во страниц, логарифмическая шкала)

Файловые онлайн-архивы и Google Scholar

Развитие парадигмы открытого доступа воплощается в создании и развитии файловых онлайн-архивов, содержащих полнотекстовые документы, как правило, в формате PDF. Данные поисковой системы Google о количестве полнотекстовых документов приведены на рис. 4.

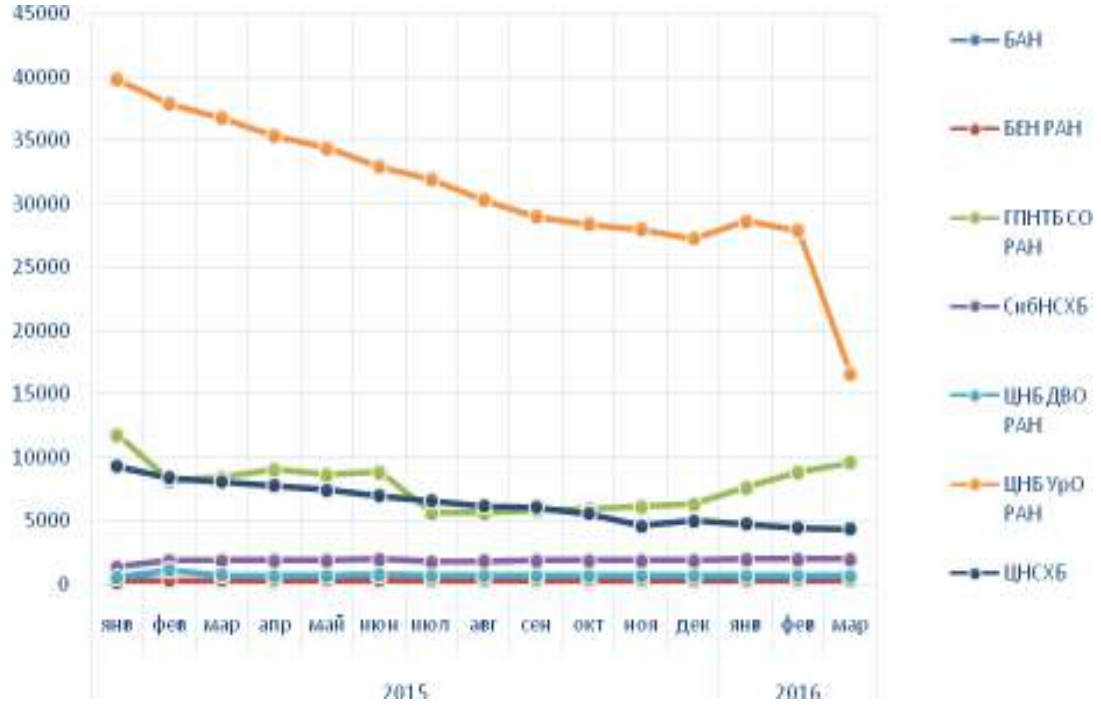

Рисунок 4. Динамика количества полнотекстовых документов на веб-сайтах по данным Google

Однако после этого падения можно увидеть разнонаправленные тренды: ГПНТБ СО РАН наращивает файловый архив (с 5630 до 9610), в ЦНБ УрО РАН и ЦНСХБ объемы сократились примерно в 2 раза, а в остальных библиотеках ситуация остается без заметных изменений. К сожалению, технические возможности на текущий момент не позволяют провести более детальный анализ, чтобы понять за счет каких именно ресурсов и разделов сайта происходят эти изменения.

Еще одним авторитетным вебометрическим показателем является количество научных публикаций, размещенных на веб-сайте и проиндексированных системой Google Scholar. Причем, решающим фактором отбора является корректное описание метаданных публикаций, а наличие полного текста не является обязательным. Этот показатель является достаточно стабильным и надежным по сравнению с другими вебометрическими индикаторами.

На рис. 5 видно несколько резких спадов (они могут быть связаны как с особенностью индексирования, так и с реальным отключением некоторых ресурсов) и отдельные сегменты умеренного роста (что свидетельствует о вероятном развитии ресурсов). Общий вывод, который следует из этих рисунков, состоит в том, что в библиотеках недостаточно системно проводится работа по развитию электронных научных библиотек открытого доступа.

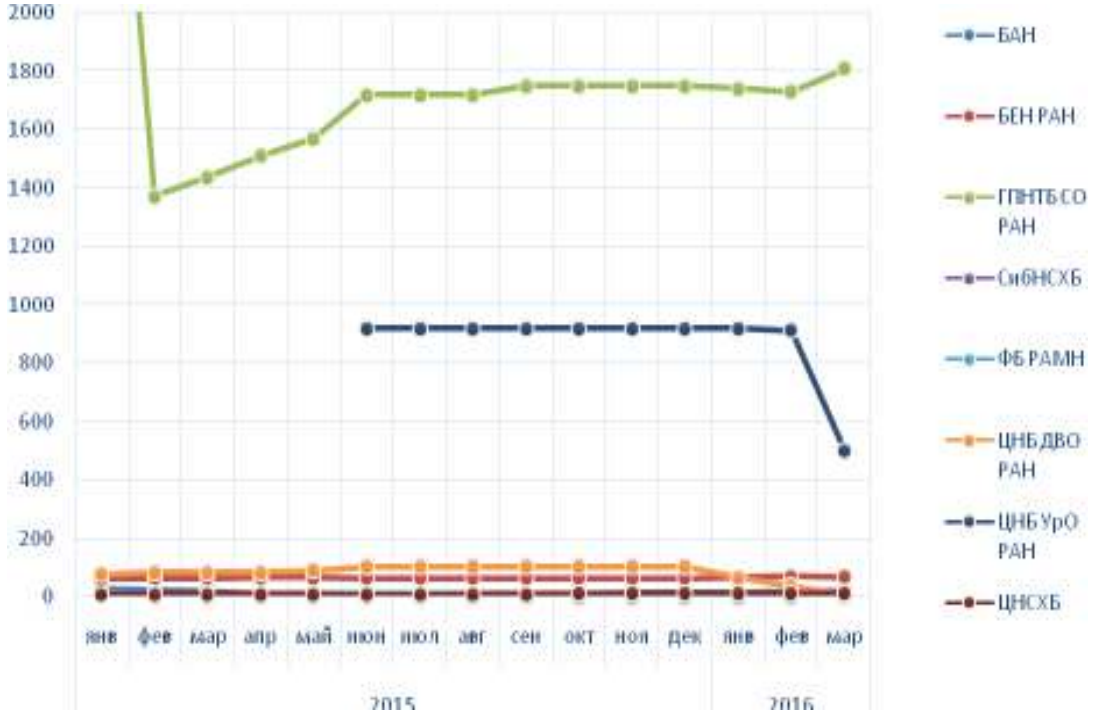

Рисунок 5. Динамика количества публикаций, проиндексированных в системе Google Scholar

Вклад библиотек в академическое веб-пространство

Отдельный интерес представляет собой сопоставление вебометрических показателей сайтов библиотек и других научных организаций. Для сравнения мы посчитали среднее значение каждого показателя для библиотек и для научных институтов, входящих в РАН (до реформы). Мы не стали учитывать данные по институтам РАМН и РАСХН, поскольку их веб-сайты в среднем развиты заметно хуже.

Из рис. 6 видно, что библиотеки имеет заметно большее количество страниц (Google-Pages – в 20 раз, Yandex-pages – в 6 раз) за счет электронных каталогов и других ресурсов. По объему полнотекстовых ресурсов преимущество также есть, хотя оно уже не такое больше (Google-Files – в 7 раз, Yandex-Files – в 2 раза), а среднее количество научных публикаций (Scholar-Papers) на сайтах библиотек сопоставимо с средним веб-сайтом института.

Среднее количество внешних ссылок на сайты библиотек (Ahref-Backlinks, Majestic-Backlinks) и среднее количество ссылающихся доменов (Ahref-Domains, Majestic-Domains) для библиотек и институтов. И это плохой показатель, т.к. уже было показано что количество страниц на сайтах библиотек многократно выше. И это означает что удельное количество внешних ссылок на одну страницу библиотеки многократно ниже, чем на веб-страницу института. Иными словами, веб-страницы библиотек являются менее авторитетными. При этом тематический индекс цитирования Яндекса (Yandex-TCY) в среднем оценивает библиотеки в 3,5 раза выше, чем институты. Соответственно, ведущие библиотеки (ЦНСХБ, ГПНТБ СО РАН, ЦНБ ДВО РАН, БЕН РАН) по размеру сайтов попадают в первую двадцатку среди организаций ФАНО. По количеству файлов в первой двадцатке бывают уже только ЦНСХБ и ГПНТБ СО РАН, а по количеству документов в Google Scholar – только ГПНТБ СО РАН.

При этом, научно-исследовательские институты совокупно устойчиво приращивают объемы файлов и индексируемых Google Scholar документов на 10-20% в год, при явном отсутствии соответствующей динамики в библиотечном сегменте.

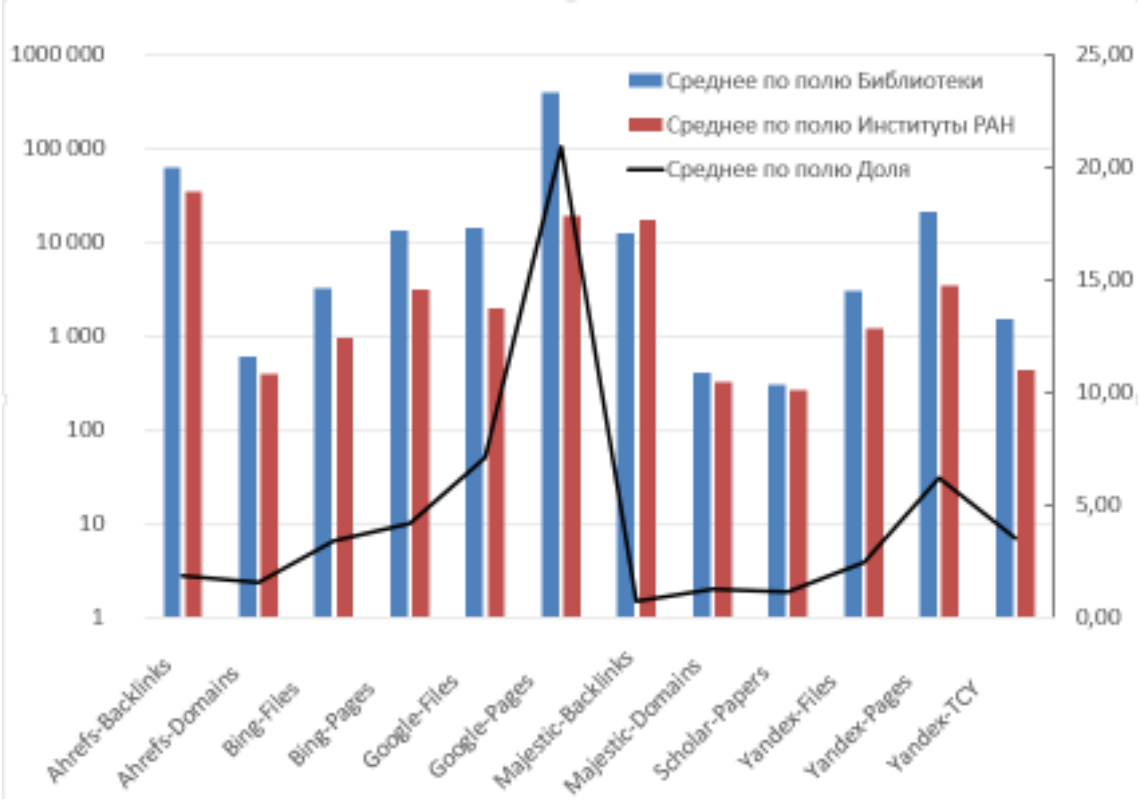

Рисунок 6. Сопоставление средних показателей сайтов библиотек и институтов РАН (до реформы)

Сравнение с другими библиотеками

Мы провели анализ вебометрических показателей крупных библиотек России, а также наиболее известных проектов электронных библиотек (см. табл. 1, данные получены в марте 2016 года). Для сравнения из группы академических библиотек мы добавили ГПНТБ СО РАН, которая по большинству показателей входила в лидирующую группу.

Таблица 1. Сравнение вебометрических характеристик крупных библиотек

Ресурс	Google-Pages	Google-Files	Scholar	Ahref-Backlinks
cyberleninka.ru	732 000	709 000	940 000	1 644 124
mathnet.ru	717 000	94 600	263 000	436 649
ГПНТБ СО РАН	494 400	14 150	3 690	435 399
НЕИКОН	240 000	13 100	91 200	3 473 398
ГПНТБ России	397 900	6 390	4 570	3 946 156
РНБ	774 000	4 180	142	3 274 164
РГБ	529 060	3 347	101	5 529 955
socionet.ru	376 000	2 150	447	99 124

Из этой таблицы видно, что количество веб-страниц отличается не более чем в 3 раза, а по остальным показателям разница может составлять 1-2 порядка и более. Безусловным лидером в области открытого доступа является проект cyberleninka.ru, второе место уверенно держит онлайн-архив mathnet.ru. Следует отметить Архив научных журналов, созданный компаний НЭИКОН, хорошо проиндексированный в Google Scholar. А наиболее авторитетными ресурсами

по количеству внешних ссылок обладают РГБ, ГПНТБ России, НЭИКОН и РНБ. Что касается ГПНТБ СО РАН, то в этой группе ее уже нельзя отнести к лидерам.

Для сравнения на рис. 7 приводим динамику показателей Google-Files и Scholar-Papers для двух крупнейших в стране онлайн-архивов CyberLeninka и Mathnet. Первый демонстрирует уверенный рост на протяжении всего года, а второй тоже растет, но очень умеренными темпами.

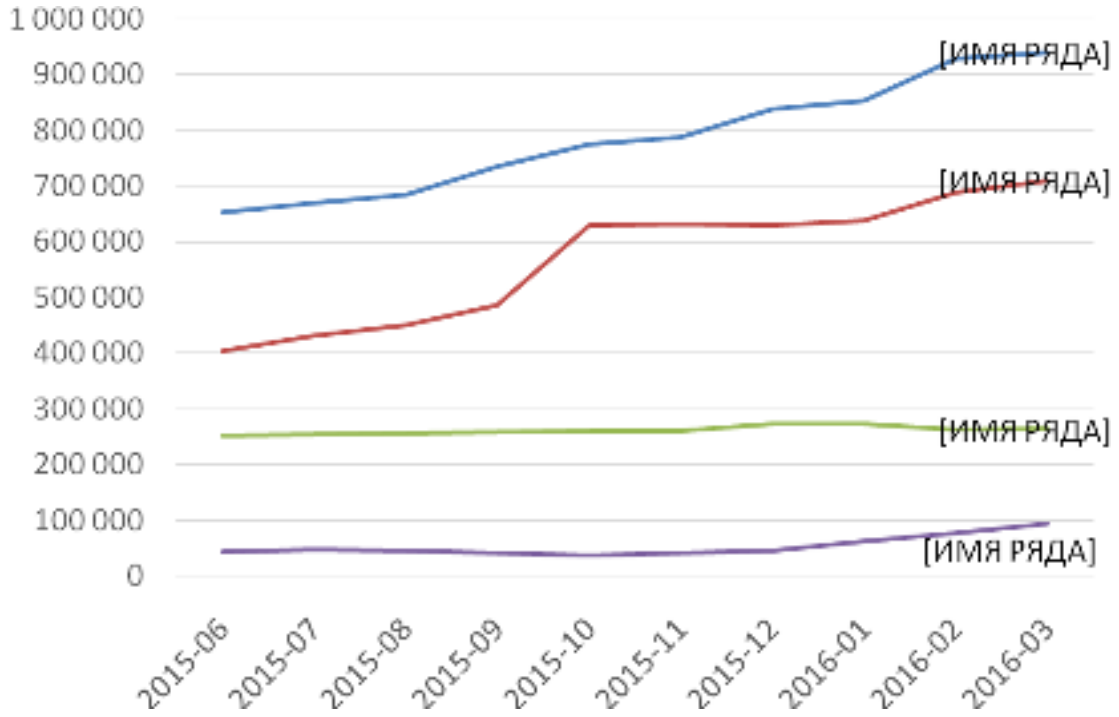

Рисунок 7. Динамика показателей Google-Files и Scholar-Papers для двух онлайн-архивов CyberLeninka и Mathnet

В итоге можно сделать вывод, что крупные библиотеки примерно одинаково относятся к развитию своих веб-ресурсов, однако отношение к отрытому доступу уже заметно отличается, а продвижение научных ресурсов в этой парадигме некоторыми библиотеками совсем не ведется.

Заключение

Исследование показало, что текущее состояние онлайн-ресурсов в академических библиотеках заметно отличается. Вебометрические индикаторы демонстрируют стагнацию, либо слабый рост, что говорит об отсутствии развития и о недостаточности усилий библиотек в этом направлении.

Причин у этого печального явления, вероятно, несколько. К ним могут относятся следующие: недостаточные знания сотрудников библиотек в области ИТ-технологий, отсутствие стратегии развития информационно-библиотечных ресурсов в онлайн среде, неприспособленности систем библиотечной автоматизации для поисковой оптимизации (SEO), нехватка финансовых и кадровых ресурсов.

Хорошим импульсом для преодоления этих трудностей может стать инициатива Федерального агентства научных организаций России по созданию единой сети информационного обеспечения научных организаций, в рамках которой разрабатывается стратегия информационного обеспечения, в т.ч. посредством развития онлайн-сервисов. Но в конечном итоге, ключи к успеху библиотек на пути в манящее цифровое будущее находятся в руках самих библиотекарей.