НАУКОМЕТРИЯ. БИБЛИОМЕТРИЯ

УДК 025.4.03

Н. А. Мазов, В. Н. Гуреев

Новые методы формирования публикационного профиля научной организации в сети науки

Представлены отработанные авторами методы сбора данных о публикациях организации, а также методы обработки профиля автора и организации во внешних БД. На основе таких выверенных данных впоследствии можно проводить качественные полномасштабные наукометрические исследования.

Ключевые слова: наукометрия, библиометрия, Web of Science, Scopus, РИНЦ, идентификация метаданных, уникальные идентификаторы публикаций.

Библиометрия и – шире – наукометрия значительно упрочили свои позиции в качестве научных дисциплин, чему способствовало развитие информационных технологий, которые сделали реферативные БД более функциональными и доступными широкому кругу пользователей. Значимость наукометрических исследований, дающих возможность моделировать научную политику, напрямую подтверждается проявленным к ним интересом со стороны не только научных организаций, но и финансирующих ведомств, а также государственных органов.

В России в этом аспекте много внимания уделяется проекту «Карта науки России» [**1**]. Отдельно упомянем разрабатываемый комплекс мер по увеличению доли отечественных публикаций в зарубежных реферативных БД [**2**]. Наряду с собственными разработками владельцев баз данных, крупнейшими из которых являются *Web of Knowledge, Scopus* и *Google Scholar*, появилось множество специализированных программ [**3**], а в недавней публикации [**4**] сделан обстоятельный обзор программного обеспечения, работающего с элементами упомянутых БД, и инструментарий этих программ позволяет проводить действительно интересные и качественные исследования по анализу науки.

Такие значимые и масштабные проекты не будут успешными, если в самом нижнем звене цепочки – на уровне конкретных научных организаций – не будет, во-первых, решена задача по сбору информации о публикациях своих научных сотрудников и, во-вторых, упорядочена информация об авторах, научных группах и организации в целом во внешних реферативных БД.

В задачу сбора информации входит создание и постоянное поддержание внутренней БД трудов сотрудников организации. Специфика решения второй задачи обусловлена различной архитектурой внешних реферативных БД. Например, в случае с *Google Академия* необходимо обеспечить доступ индексирующим роботам *Google* к своей внутренней БД. В *Web of Knowledge* и *Scopus* предполагается непосредственное участие назначенных в организации лиц в редактировании профиля организации.

В любом случае решение главных наукометрических задач должно идти снизу вверх и начинаться на уровне организации, чтобы впоследствии можно было легко объединять данные в более общие большие кластеры – по ведомственной принадлежности, географическим признакам, предметным областям и др.

С большой долей уверенности можно сказать, что в настоящее время работа по обоим обозначенным направлениям во многих научных организациях ведётся неудовлетворительно, что обусловлено рядом обстоятельств. Прежде всего, остаётся нерешённым кадровый вопрос. Поскольку библиометрия – наука молодая, во многих институтах ещё не определились с тем, кто должен ею заниматься – учёные или секретариат, сотрудники библиотеки или специально созданных подразделений. Во-вторых, отсутствуют хорошие методики проведения такого комплекса работ. В-третьих, спрос на подобные работы возрос лишь в последнее время.

Оставляя в стороне вопросы организации процесса, мы предлагаем вниманию читателей ряд методологических рекомендаций, основанных на опыте многолетней работы двух научных институтов – Института нефтегазовой геологии и геофизики им. акад. Трофимука (ИНГГ) СО РАН и Государственного

научного центра вирусологии и биотехнологии (ГНЦ ВБ) «Вектор».

Прежде всего стоит отметить, что ситуация в российских научных организациях не уникальна – подобные проблемы свойственны и многим зарубежным институтам. Лишь в нескольких крупных организациях работа проводится на качественном уровне. Информационные центры, достигшие успеха в этом направлении, указывают на необходимость задействовать ручной труд в процессе сбора библиометрической информации и на невозможность использования лишь внешних реферативных БД [**5**, **6**].

Если говорить о российских НИИ, то здесь проблема усугубляется различными способами транслитерации русских имён, названий организаций и журналов, заглавий статей (которые могут не только транслитерироваться, но и переводиться), а также тем, что в качестве места проведения работы указывается не свой институт (головная организация), а сторонняя организация, где автор работает по совместительству, или же сведения об организации отсутствуют полностью.

На трудности обработки именно кириллической информации в зарубежных БД указывают сами зарубежные исследователи [**7**] (заметим: такие же трудности возникают при обработке иероглифических библиографических сведений). Отмеченные проблемные моменты приводят к тому, что в зарубежных БД множатся профили одного и того же автора, одной и той же организации, сведения об одной и той же публикации. Особую трудность в данном случае представляет отсутствие единых стандартов на идентификацию метаданных [**8**, **9**].

В такой ситуации остаётся либо ожидать разработки уникальной идентификации записей в БД крупными игроками издательского дела и разработчиками стандартов, либо использовать доступный инструментарий, предоставляемый БД, а также свои собственные методы.

На базе ИНГГ СО РАН и ГНЦ ВБ «Вектор» созданы и отработаны полуавтоматическая система сбора информации о публикациях, а также методы обработки информации во внешних реферативных БД.

Ведение базы данных публикаций. Непременное условие успешной работы – сбор информации обо всех публикациях организации со времени её создания. Эта работа трудоёмкая и может занять до нескольких месяцев, однако её необходимо провести лишь один раз.

Учитывая, что большинство наукометрических исследований основано на списках публикаций из БД *Web of Knowledge* или *Scopus*, нужно обязательно проверить, имеется или отсутствует полный список всех публикаций организации в этих БД. Затем следует выписать все идентификаторы статей, а также авторские идентификаторы и сохранить их в любой локально используемой БД или текстовом редакторе. После этого мы получим возможность ввести идентификаторы в качестве запроса в интересующую нас БД и получить результат – полный список публикаций организации на текущий момент.

Следующий шаг направлен на поддержание списка в актуальном состоянии, для чего создаётся рассылка (*Alert*) для каждого автора, работающего в организации. При индексировании статьи базой данных информация о ней направляется по электронной почте сотруднику, ответственному за ведение БД, и он добавляет идентификатор статьи в общий список идентификаторов организации.

Мы акцентируем внимание на отработке именно полного списка статей при работе с любой БД, поскольку в публикации авторы зачастую либо вообще не указывают головную организацию, либо указывают другую, в которой также числятся. Поэтому запрос по адресу организации, как мы уже отметили, зачастую выдаёт лишь половину результатов.

Описанные алгоритмы изучения публикационной активности разработаны и успешно применяются в информационно-библиотечном центре ИНГГ СО РАН на протяжении последних 15 лет.

Информационно-библиотечная система НТИ по наукам о Земле ИНГГ СО РАН – одна из наиболее совершенных не только в Сибирском отделении РАН, но и в России. Сегодня информационные массивы содержат десятки миллионов библиографических записей и сотни тысяч полнотекстовых документов.

Создана БД «Труды сотрудников ИНГГ и ИГМ СО РАН», где представлены библиографические описания публикаций, которые имеются в фондах библиотеки института и представлены на его сайте. БД создавалась в среде АБИС CDS/ISIS; включает описания монографий, диссертаций, авторефератов диссертаций, статей в

научных журналах, электронных публикаций в Интернете, докладов на конференциях и других публикаций.

Для интеграции описаний с БД *Web of Science*, *Scopus* и РИНЦ разработан программно-технологический комплекс, позволяющий снабжать метаданные соответствующими ссылками. К описаниям документов, отражённых в перечисленных БД, добавляются соответствующие идентификаторы записи, однозначно указывающие на соответствующую запись в этих БД. Такая технология доработки записей позволяет в оперативном режиме отслеживать публикуемость и цитируемость работ в перечисленных БД.

В настоящее время БД «Труды…» содержит около 40 тыс. записей. Для более 12 % описаний имеются ссылки на электронные версии публикаций. БД размещена на сайте ИНГГ СО РАН и доступна по адресу http://ibc.ipgg.nsc.ru.

Представленный опыт в настоящее время успешно прошёл тестирование в ГНЦ ВБ «Вектор».

Редактирование профилей авторов, научных групп и организаций во внешних реферативных БД. Сегодня мы наблюдаем заметный прогресс в области разработки уникальных идентификаторов в цифровой медиасреде. Если прежде все усилия владельцев крупных БД были направлены на унификацию метаданных лишь на своем ресурсе, то 2012 г. был отмечен договорами между разными разработчиками в области унификации авторских профилей. Так, сейчас стало возможным указывать в определённой системе свои уникальные идентификаторы из других БД. Например, в *Web of Science* можно указать *ORCID* (разработка Эльзевир), а в РИНЦ – и *ORCID,* и *Researcher ID* от *Web of Science*.

В наших институтах в своё время были отработаны все авторские профили в БД *Scopus* и проведён сравнительный анализ возможностей *WoS, Scopus* и РИНЦ, предоставляемых представителям организации и авторам публикаций [**10**]. В настоящее время тестируется разработанная РИНЦ система *Science Index – организация.*

Во всех реферативных БД при обработке информации о метаданных используется большое количество параметров, в том числе: общедоступная информация об авторе в Интернете, цитирования и самоцитирования, информация с первой страницы публикации и др. Тем не менее лучшими разработками являются полуавтоматические, требующие участия человека в процессе обработки данных [**11**]. Именно поэтому разработчики БД для повышения точности информации об авторах и организациях всё чаще делегируют им права на редактирование данных о своих публикациях.

Редактирование может быть прямым, как в *Web of Science* или *Google Академия*, где автор может вручную внести свои работы в общий список. Однако такие изменения сохраняются лишь в авторском профиле, который хранится отдельно от основных БД. Следовательно, авторские правки не отражаются в основном массиве данных.

Редактирование может быть опосредованным, как в БД *Scopus*, где автор отправляет запрос техническому персоналу, который проверяет данные на достоверность и вносит изменения в основную БД. Однако в этом случае автор не может добавить в свой профиль работы, не расписываемые в *Scopus*.

В РИНЦ разработчики нашли свой оригинальный подход: авторы могут работать только с публикациями, проиндексированными системой, но все их операции напрямую отражаются в основной БД. Мы заметили, что этим вполне могут воспользоваться недобросовестные пользователи, причислив работы (и цитирования) однофамильцев к своему профилю.

Редактирование профиля организации проводится авторизованным представителем и соответствует алгоритмам, разработанным в базах данных для авторов. Так, в *Scopus* работа проводится путём запросов к технической поддержке, сотрудники которой проверяют информацию от авторизованного представителя и вносят изменения в БД.

В РИНЦ вся ответственность ложится на сотрудника той или иной организации, который без стороннего контроля вносит изменения в основную БД. Как и в случае с авторскими профилями, здесь кроется много возможностей для злоупотреблений, когда организации приписываются работы, проведённые в других учреждениях. Заметим, что такие случаи уже есть.

В *Google Академия* нет инструментов, аналогичных другим БД. Между тем представитель организации

может выставить для индексирования роботами *Google* данные своей внутренней БД в статических html-страничках, снабдив соответствующие метатэги необходимой информацией о публикации, согласно описанию метаданных Дублинского ядра.

Наиболее сбалансированный подход представлен в БД *Scopus*, в которой ни авторам, ни представителям организации напрямую не предоставлены права на внесение изменений, но всё продумано для того, чтобы и автору, и представителю было легко сообщить о неточностях и внести нужные коррективы. Существенно то, что внесенные изменения проверяются и впоследствии отражаются в самой БД, а не в отдельной надстройке, что делает эту БД более точной и удобной для поиска информации. Итогом совместной корректировки сотрудниками той или иной организации и модераторами списка публикаций становится выверенный профиль автора и организации.

В заключение отметим: рассмотренная методика обработки с января 2012 г. находится в промышленной эксплуатации в ИНГГ СО РАН. Данные, получаемые из *WoS*, *Scopus* и РИНЦ, полностью интегрированы в собственную БД «Труды сотрудников ИНГГ и ИГМ СО РАН». А работа с БД *Scopus* прошла апробацию на полномасштабной отработке профиля организации в ГНЦ ВБ «Вектор» и ИНГГ СО РАН.

Список источников

1. **Проект** создания «Карты Науки России». Установочное заседание экспертных групп проекта : материалы для обсуждения. – 2013 [Электронный ресурс]. – Режим доступа: http://www.lin.irk.ru/new/files/sci_ross.pdf (Дата обращения: 29.04.2013).

2. **Проект** распоряжения Правительства РФ «Комплекс мероприятий, направленных на увеличение к 2015 году доли публикаций российских исследователей в общем количестве публикаций в мировых научных журналах, индексируемых в базе данных «Сеть науки» (Web of science) до 2,44%». – 2013 [Электронный ресурс]. – Режим доступа: http://минобрнауки.рф/РґРѕРєСѓРјРµРЅС‚С‹/3119/С„Р°Р№Р»/1847/13.02.27-СЃРµС‚СЊ.PSP°СѓРєРё.pdf (Дата обращения: 29.04.2013)

3. **Мазов Н. А**. Свободно распространяемые программы для наукометрических и библиометрических исследований // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: 19-я междунар. конф. "Крым–2012" (2–10 июня 2012 г., г. Судак) : труды конф. – Москва : ГПНТБ России, 2012. – С. 1–6. [Электронный ресурс]. – Режим доступа: http://www.gpntb.ru/win/inter-events/crimea2012/disk/123.pdf (Дата обращения: 29.04.2013)

4. **Cobo M. J., López-Herrera A. G., Herrera-Viedma E., Herrera F**. Science Mapping Software Tools: Review, Analysis, and Cooperative Study Among Tools // Journal of the American Society for Information Science And Technology. – 2011. – V. 62(7). – P. 1382–1402.

5. **Raan A. F. J. van**. The use of bibliometric analysis in research performance assessment and monitoring of interdisciplinary scientific developments // Technikfolgenabschätzung – Theorie und Praxis. – 2003. – V.1(12). – P. 20–29.

6. **Bibliometrics.** Publication Analysis as a Tool for Science Mapping and Research Assessment. The Karolinska Institutet Bibliometrics Project Group. – 2008. [Электронный ресурс]. – Режим доступа: http://ki.se/content/1/c6/01/79/31/introduction_to_bibliometrics_v1.3.pdf (Датаобращения: 29.04.2013).

7. **Egghe L., Rousseau R.** Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science. – Amsterdam : Elsevier science publishers, 1990. – P. 217–218.

8. **Vitiello G.** Identifiers and Identification Systems: An Informational Look at Policies and Roles from a Library Perspective // D-Lib Magazine. – 2004. – V. 10(1) [Электронный ресурс]. – Режим доступа: http://www.dlib.org/dlib/january04/vitiello/01vitiello.html (Дата обращения: 29.04.2013).

9. **Мазов Н. А., Гуреев В. Н., Жижимов О. Л.** Единая идентификация библиографических метаданных: проблемы и решения. Распределенные информационные и вычислительные ресурсы (DICR-2012) : Материалы XIV Рос. конф. с участием иностр. учёных (26–30 нояб. 2012 г., Новосибирск). – Новосибирск :

ИВТ СО РАН. – 2012. – С. 18.

10. **Мазов Н. А., Гуреев В. Н**. Новые научные методы для исследования библиотечной отрасли // Библиосфера. – 2012. – № 5. – С. 87–90.

11. **Kang I.-N., Kim P., Lee S., et al.** Construction of a large-scale test set for author disambiguation // Information processing and management. – 2011. – № 47. – P. 452–465.