ЧТО ДЕЛАТЬ С БОЛЕЗНЯМИ БИБЛИОГРАФИЧЕСКИХ БАЗ

Т.А. Полилова

Институт прикладной математики им. М.В. Келдыша РАН

Сбор информации о публикационной активности становится последнее время регулярной задачей, требующей немалых усилий со стороны ученых и научных организаций. Запрашивающие органы часто интересует не только простой перечень публикаций, но и такие характеристики, как присутствие статей в известных библиографических базах, цитируемость статей, импакт-факторы изданий, привязка статей к тематике или специальностям ВАК и т.д.

Существует множество зарубежных и российских библиографических баз и проектов, занимающихся сбором информации о научных публикациях. Базы располагают достаточно развитым аппаратом поиска статей, однако автору и организации не всегда удается отобрать интересующие статья «в один клик». Чаще всего приходится отбирать статьи в более широком поле результатов запросов, с помощью нескольких запросов с вариантами поисковых параметров, или с помощью хитроумных комбинаций запросов и последующей ручной работой по отбору нужных данных.

Библиографические базы, такие как Web of Science, Scopus, eLibrary и др., формируются на основе поставляемой издателями информации. Издатели, как показывает практика, часто не в состоянии предоставить точную информацию об авторах и их организациях. Уже на этом уровне возникает первая волна проблем библиографических баз — затруднения в идентификации автора, в привязке автора к научной организации (месту его работы).

Для получения показателей цитируемости библиографические базы выполняют разбор библиографических ссылок, содержащихся в тексте статей. Далее делается попытка отождествить имеющиеся в статьях ссылки с хранящимися в базе записями, относящимися к цитируемым статьям. При благоприятном стечении обстоятельств цитируемые статьи и их авторы увеличивают свои показатели цитируемости. Но поскольку в статьях содержится значительное число неполных или искаженных библиографических ссылок, немалая часть цитирований в библиографических базах не учитывается. Это обстоятельство вызывает вторую волну проблем — ошибки в подсчете цитирований статей, которая в свою очередь порождает третью волну — ошибки в показателях цитируемости (индекс Хирша авторов, импакт-фактор изданий).

В качестве примера сошлемся на собственный опыт получения данных по цитированию научного издания «Препринты ИПМ им. М.В.Келдыша» в Российском индексе научного цитирования (РИНЦ) на базе eLibrary. В таблице eLibrary у издания показатели цитируемости были сформированы на основе учета «правильных» библиографических ссылок. К сожалению, при цитировании препринта авторы часто искажали название издание (мы зафиксировали несколько десятков вариантов искажения названия), и такие цитирования не учитывались. В результате ручного отбора ссылающихся на препринты статей удалось поднять более чем втрое импакт-фактор издания, подсчитанный аппаратом eLibrary.

Самой точной информацией о своих публикациях располагает автор, аккуратно ведущий список научных трудов. Результаты обращений к библиографическим базам чаще всего дают лишь приблизительные показатели публикационной активности автора. Эта очевидная проблема во многих базах решается с помощью механизма, позволяющего автору вручную «привязать» нераспознанные в системе публикации и уточнить сведения о своих публикациях и цитированиях. Таким образом, сейчас ученому фактически предложено стать контент-менеджером многочисленных баз и проектов.

Существенно более экономичным, рациональным и комфортным, на наш взгляд, было бы решение — заниматься лишь очевидными насущными задачами: в удобной среде создавать и регулярно обновлять свой единственный список научных трудов, а далее автоматически получать совокупные достоверные сведения о цитированиях и сопутствующие «индексы». Однако существующие библиографические базы и аналогичные им проекты не готовы к подобной интеграции и обмену данными.

Отметим, что далеко не все научные труды попадают в библиографические базы. Поставщиками данных могли бы служить сайты научных организаций, где представлены научные результаты сотрудников [1]. Сайты научных организаций отличаются большим разнообразием, и, к сожалению, размещенная там информация сейчас практически непригодна для включения во внешние инфраструктурные проекты.

Научной общественности хотелось бы получить современные технологические решения, которые, с одной стороны, обеспечили простой и удобный способ ведения списка научных трудов ученого, с другой стороны, экранировали ученого и научную организацию от многочисленных кампаний по сбору сведений о публикационной активности: такие сведения заинтересованные структуры могли бы получать самостоятельно путем формальных запросов к формируемым таким образом базам.

Что можно было бы реализовать в существующих условиях? Рациональным и жизнеспособным представляется механизм, с помощью которого автор размещает информацию о себе и своих результатах лишь в одном месте, например, на персональной странице на сайте научной организации, дополняя эту информацию метаданными, доступными для анализа, пополнения и коррекции заинтересованными хранилищами данных. Состав метаданных о научных результатах в большинстве случаев достаточно очевиден. В частности, рациональный состав атрибутов научной публикации можно сформировать, проанализировав наиболее известные проекты в этой сфере: BibTeX, Dublin Core, Google Scholar, schema.org, Открытые архивы (OAI-PMH), BIBFRAME и др.

Прообразами такого механизма можно считать известный протокол OpenID, число пользователей которого перевалило за миллиард, и родственный проект ORCID, нацеленный именно на научные приложения. Здесь пользователь, единожды сообщивший о себе некоторую совокупность сведений, получает возможность разрешать передавать эти сведения вовне, избегая тем самым многократных повторений рутинных действий, требующихся при заполнении разнообразных форм в многочисленных библиографических базах. Однако в этих проектах основным и по существу единственным обслуживаемым объектом является отдельный человек.

Нас же интересует интегральное информационное обслуживание научной организации и ее сотрудников. Включение в рассмотрение не только сотрудника, но и организации несет множество преимуществ. В частности, такие формальные сведения о сотруднике, как его должность, ученая степень, государственные награды, более естественно и надежно поддерживаются отделом кадров организации. На этом принципе основана система создания и обслуживания персональных страниц сотрудников ИПМ им. М.В. Келдыша РАН: формальные сведения попадают на персональные страницы через отдел кадров, где посредством соответствующего приложения формируется строго официальная часть персональных страниц сотрудников [2].

На персональной странице научного сотрудника должна аккумулироваться информация о результатах его научной деятельности (сведения о публикациях, отчетах, патентах и пр.). Часть таких сведений может быть представлена в форме соответствующих обращений к библиографическим базам.

Хранимые в системе данные должны быть доступны извне. Например, научный сотрудник может разрешить конкретной библиографической базе на постоянной основе использовать хранящуюся на персональной странице информацию о нем. Или же организация в целом, вместо того чтобы составлять очередной затребованный отчет, открывает контролирующей структуре доступ к данным своих сотрудников. Такой подход обеспечивает эволюционное развитие научного пространства «снизу-вверх». Существующие технологии интернета позволяют реализовать этот подход уже сегодня.

Одним из перспективных способов представления сведений о научных результатах автора на его персональной странице является аппарат микроформатов, в частности — микроданные HTML5. Аппарат микроформатов позволяет превратить сайт в объект для автоматической семантической обработки. Метаданные веб-страниц, представленные микроформатами, становятся доступными для краулера (робота) любого заинтересованного проекта, собирающего информацию в согласованном формате.

Важно и другое обстоятельство. Благодаря предлагаемой технике данные о результатах нескольких организаций легко объединяются в единую семантическую сеть, формируются горизонтальные связи между учеными. В этой комфортной среде научному сообществу становится доступной самая полная, надежная и достоверная профессиональная информация. Когда научная статья вместе с набором метаданных размещается в интернете, она становится новым элементом семантически связанного пространства научных публикаций.

А ученый становится в предлагаемой среде основным поставщиком первичных данных и заинтересованным участником формирующейся научной коммуникации.

В заключение кратко перечислим основные болезни библиографических баз: неточный подсчет числа библиографических ссылок на публикацию, как следствие – искажение показателей цитируемости (индекса Хирша авторов, импакт-фактора изданий). Из-за ошибок при разборе слабо формализованных библиографических ссылок показатели цитируемости нередко занижаются в несколько раз, что дискредитирует технологию библиографических баз. В качестве альтернативы предлагается развивать такие технологии, где поставщиками первичной информации являются автор, научная организация, и где содержательные связи реализуются более надежными механизмами, доступными для автоматической семантической обработки. Примеры таких технологий – проект ORCID, аппарат микроданных HTML5.

Работа выполнена при поддержке РФФИ в рамках гранта № 13-01-00493 а.

ЛИТЕРАТУРА:

1. Т.А. Полилова. Инфраструктура научных публикаций // Международный форум по информации. М: ВИНИТИ РАН. 2009. Т. 34. № 3. С. 3-12.
2. Т.А. Полилова. Персональные веб-страницы в научном сообществе. Труды Международной конференции "Научный сервис в сети Интернет: экзафлопсное будущее" (19-2 сентября 2011 г., г. Новороссийск). Элект. издание, 2011. С. 476-479. ISBN 978-5-211-06229-0.