О качестве контента в интегрированных системах на примере Карты российской науки

Content quality of integrated systems as exemplified by the Map of Russian Science OntIntegrated syste

И. В. Михайленко, Т. В. Лясникова, Е. М. Гончарова
Государственная публичная научно-техническая библиотека России,
Москва, Россия

Irina Mikhailenko, Tatyana Lyasnikova and Elena Goncharova
Russian National Public Library for Science and Technology,
Moscow, Russia

Доклад содержит описание типичных ошибок, обнаруженных при работе с контентом в интегрированной информационно-аналитической системе агрегации данных о цитированиях «Карта российской науки». Произведен анализ причин возникновения ошибок на различных этапах передачи данных от создателя публикации – к системе агрегации данных об индексах цитирования.

Described are typical mistakes of working with the content of the integrated analytical citation data aggregation system Map of Russian Science. The causes for mistakes at different levels of data transfer from publication originators to the system aggregating data on citation indices are analyzed.

Проект «Карта российской науки» (КРН) исходно ориентирован на интеграцию информационных ресурсов крупнейших систем научного цитирования, а именно Web of Science (WoS), Российский индекс научного цитирования (РИНЦ) и др. Безусловно, каждая из этих систем имеет свою специфику импорта и обработки данных, свои подходы к классификационным системам, свои поисковые системы, транслитерации имен, фамилий. При обработке конкретных запросов со стороны ученых и специалистов к КРН такая специфика систем цитирования порождает целый ряд вопросов и трудностей при обработке данных.

Некоторым причинам возникающих сложностей посвящена настоящая работа.

В условиях необходимости работы с несколькими различными системами цитирования, нас заинтересовал вопрос о том, насколько абсолютны системы такого рода. Мы задались целью выяснить, могут ли такие системы быть безошибочными, проследив весь путь данных от создания публикации до отражения публикации в наукометрических показателях автора. Можно предположить, что исследование этого вопроса важно не только специалистам по работе с наукометрией, но и конечным пользователям таких данных – ученым, руководителям научных подразделений организаций, а так же лицам, участвующим в процессе принятия решений в сфере управления наукой.

Для ответа на заданный вопрос, мы хотим подробнее рассмотреть источники данных наукометрических систем, процессы импорта этих данных в системы, а так же ошибки, возникающие при произведении технологических операций.

Для начала выявим все технологические операции процесса получения наукометрических показателей (рис.1). Первым этапом является создание статьи (книги) автором (этап 1). Затем рукопись передается в издательство и проходит редакторскую правку (этап 2). После этого издательство передает электронную (этап 3a) или печатную копию (этап 3b) публикации в организацию-создателя наукометрических данных. Далее возможно два варианта работы – в зависимости от формата носителя передаваемых данных. Данные либо пересылаются в электронной форме по сети интернет (этап 4a), либо переводятся организацией-получателем в электронный вид (этап 4b). На следующем этапе происходит предварительная обработка входных данных – перевод в необходимый формат, установка связей между объектами (этап 5). Затем обработанные данные импортируются в информационную систему (этап 6). Только после этого система рассчитывает собственно наукометрические индексы (этап 7).

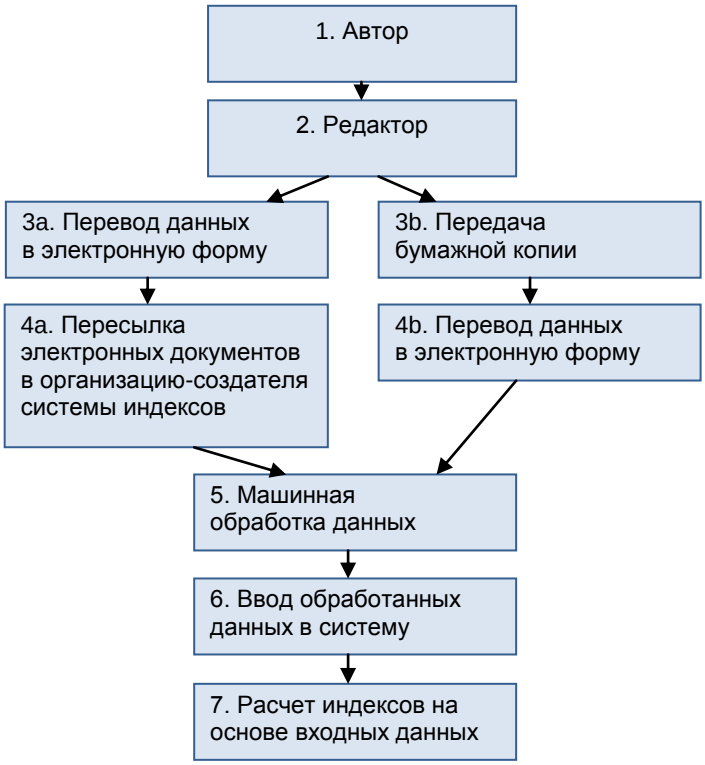

Рис. 1. Процесс создания наукометрических индексов

На каждом этапе могут возникать различные типы ошибок. Рассмотрим типы вероятных ошибок подробнее.

Первый тип ошибок – опечатки, он возникает на тех этапах обработки данных, где участвует человек (1 и 2 этапы, рис.1). Про опечатки вышла целая книга (3), в которой рассказывается об опечатках, как в книгах, так и в статьях, в том числе и научных. Ошибки в написании фамилии или организации могут повлиять на индексы публикационной активности. Можно привести такой пример – в публикации (7) один из авторов – Матвеичев Алексей Валерьевич, к.ф.-м.н., сотрудник Института проблем химической физики РАН, в описании статьи в РИНЦ указан как «- Матвеичев А.В.». Операторы системы РИНЦ вручную добавили эту публикацию в список ученого, и ни она, ни её цитирование не пропало. Однако, опечатка повлекла за собой другие типы ошибок – в КРН публикация попала с таким же написанием автора, и, конечно, без связи с цитирующей её статьей. После обращения к нам мы сможем включить статью в список публикаций Алексея Валерьевича, но восстановить связь с цитирующей статьей уже будем не вправе – по лицензионному соглашению с РИНЦ, ограничивающему наши права в редактировании данных. В результате – в КРН отсутствует ссылка, индекс цитирования работ ученого занижен – по причине одной опечатки. Несмотря на урон, приносящий научной репутации, ошибки в показателях систем цитирования – не самые страшные. Встречаются и более серьёзные опечатки – такие как искажение данных в медицинских журналах (4), или в финансовых материалах (5). Если опечатка допущена в фамилии автора – в результирующих данных будет невозможно связать реального автора и его публикацию, а значит, публикация не отразится в индексах. Опечатки в списках литературы могут привести к появлению несуществующих публикаций, то есть отдельных объектов базы данных. В то же время действительная ссылка в системе будет отсутствовать, и автор, знающий, кто на него ссылался, обнаружит отсутствие ссылки и отсутствие показателей цитируемости своих работ (6). Это своеобразный обман (неумышленный) пользователей информационных систем авторами или издателями. Однако, обман может быть и умышленным: так, многие работы в системе РИНЦ из-за различий в библиографических описаниях в списки публикаций попадают по два-три раза(1), таким образом, завышая количество статей у отдельных авторов. Исключить или уменьшить долю

таких ошибок с точки зрения информационной системы достаточно сложно, поскольку это требует знания того, где и как может ошибиться человек.

Второй тип ошибок – это ошибки, возникающие при транслитерации иноязычных фамилий. Здесь можно привести множество примеров – Дударёнок Анна Сергеевна, научный сотрудник Института оптики атмосферы им. акад. В.Е. Зуева СО РАН, чью фамилию пишут и как Dudaryonok, и Dudarenok. Другой пример – Объедков Сергей Александрович, к.т.н., доцент Высшей школы экономики, в РИНЦ имеет 0 статей, по причине того что его фамилию пишут как Obiedkov, Ob'edkov, Obedkov. В результате – «0 статей» и премия Scopus Award Russia 2013 по результатам публикационной активности и цитируемости в мире. Как следствие – в КРН нет вообще профиля Сергея Александровича на русском языке, и несколько разных профилей на различные варианты написания фамилии, конечно, с потерянными ссылками цитирований.

Третий тип ошибок – это ошибки, возникающие при машинной обработке данных (этапы 3a, 4b, 5). Здесь объединены ошибки в ходе распознавания сканированных документов компьютерными программами и ошибки автоматизированного приведения данных в заданный формат (выборка необходимых данных из массива по формальным критериям). Этого типа ошибок можно было бы избежать при ручной проверке данных, однако на обрабатываемых наукометрическими системами массивах информации это невозможно в силу объема данных. Например, РИНЦ официально сообщает, что Бюджет проекта не позволяет проводить весь цикл обработки поступающей информации операторами в ручном режиме. Поэтому такие операции, как разбор ссылок или привязка публикаций и ссылок к авторам, организациям и журналам производится в РИНЦ в автоматическом режиме. Естественно, что далеко не все записи удается точно разобрать, особенно учитывая низкую культуру оформления списков цитируемой литературы в большинстве российских журналов (2).

К четвертому типу ошибок можно отнести потери данных, возникающие при передаче данных, как по сети интернет, так и в саму систему наукометрических индексов (этапы 4a, 6).

Пятый тип ошибок – это ошибки в коде системы, которые искажают результативные данные – количество цитирований, индекс Хирша (этап 7). Этот тип ошибок также можно свести к минимуму благодаря точности и внимательности разработчиков и тестировщиков системы.

Таким образом, мы получаем схему этапов передачи данных с возможными ошибками на каждом этапе (рис. 2).

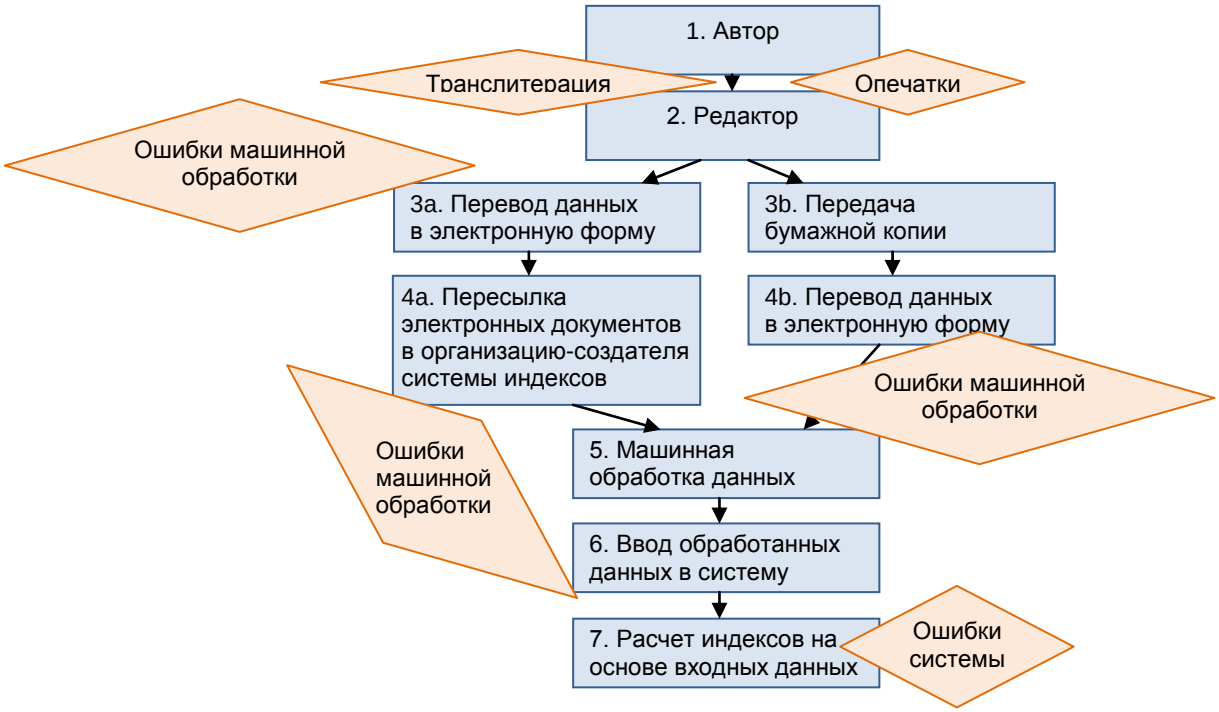

Рисунок 2. Этапы передачи данных в систему цитирования с возможными ошибками

Таким образом, можно сделать вывод о неизбежности наличия определённой погрешности при составлении наукометрических показателей.

В ходе работы с наукометрической системой Карта российской науки мы получаем запросы пользователей на исправление ошибок в данных и показателях, однако мы можем исполнить около 70-80% заявок, поскольку ошибки были допущены во внешних по отношению к системе данных. По условиям лицензионного соглашения, мы не имеем права такие данные редактировать. С аналогичными проблемами сталкивается и система Web of Science: в процессе работы мы обращаемся к их системе с заявками на исправление данных, в 15% случаев мы получаем отказ, связанный с невозможностью отредактировать внешние для их системы данные. Один из последних таких примеров – публикация к.ф.-м.н., с.н.с. Саратовского филиала Института радиотехники и электроники им. В.А. Котельникова РАН Фатеева Дениса Васильевича из материалов международной конференции (8), рецензируемых WoS. Часть авторов публикации в системе WoS указана без аффилиации с организациями. В ответ на нашу заявку мы получили уведомление о том, что данный издатель передает в систему WoS только аффилиацию первого автора, вследствие чего сотрудники WoS не могут проверить, и дополнить аффилиацию Фатеева Дениса Васильевича. В результате КРН также не получает эту аффилиацию и не вносит статью в список публикаций Дениса Васильевича и Института радиотехники и электроники им. В.А. Котельникова РАН.

Важно подчеркнуть, что часть ошибок смогут нивелировать авторы публикаций: целесообразнее потратить время и силы на дополнительную проверку статьи перед публикацией, чем на обращения к держателям различных наукометрических систем, которые далеко не всегда смогут помочь. Важнейшими пунктами проверки должны быть данные об авторе, заголовок публикации, и список использованной литературы.

Литература

1. Кузнецов, А.В. Для начала надо навести порядок в существующей системе РИНЦ /А. В. Кузнецов // Вестник российской академии наук. – 2014. – Том 84. – № 3. – С. 268–269.
2. РИНЦ и Science Index в вопросах и ответах [Электронный ресурс] / ООО Научная электронная библиотека. – Режим доступа: http://elibrary.ru/projects/science_index/science_index_questions.asp (24.04.2014).
3. Д.Ю. Шерих. «А» упало, «Б» пропало... Занимательная история опечаток [Текст]. – М.: Центрполиграф : «МиМ-Дельта», 2004. – 173 с.
4. Garcia-Berthou, E. Incongruence between test statistics and P values in medical papers [Электронный ресурс] / E. Garcia-Berthou, C. Alcaraz // BMC Medical Research Methodology. – 2004. – Том 4. – № 13. – Режим доступа: http://www.biomedcentral.com/1471-2288/4/13 (24.04.2014).
5. Цыплухин, В. Опечатка ценой 0,5 млн евро [Электронный ресурс] / E-xecutive.ru. – Режим доступа: http://www.e-xecutive.ru/knowledge/announcement/1182205/index.php?PAGE_NAME=read&FID=12&TID=8177 (24.04.2014).
6. Аникеева, О. С. Использование индекса научного цитирования в качестве характеристики научно-исследовательской деятельности ученых // Наука. Инновации. Технологии. – 2009. – №6. – С. 5–11.
7. Султанов, В.Г. FPIC3D – параллельный код для моделирования высокоэнергетических процессов в конденсированных средах / Султанов В.Г., Григорьев Д.А., Ким В.В., Ломоносов И.В., Матвеичев А.В., Острик А.В., Шутов А.В. // Вычислительные методы и программирование: новые вычислительные технологии. – 2009. – Том 10. – № 1. – С. 101–109.
8. Maremyanin, K. V. Resonance detection of terahertz radiation in nanometer field-effect transistors with two-dimensional electron gas / Maremyanin K.V., Gavrilenko V.I., Morozov S.V., Ermolaev D.M., Zemlyakov V.E., Shapoval S.Y., Fateev D.V., Popov V.V., Maleev N.A., Teppe F., Knap W. // 35th International Conference on Infrared, Millimeter and Terahertz Waves (Rome, Italy, Sep 05-10, 2010): материалы конференции. – С. 1–2.