Проблемы идентификации метаданных в наукометрических базах данных Web of Knowledge, Scopus и РИНЦ на примере профилей авторов

Problems of Identification of Metadata in Scientometric Databases WoK, Scopus and Russian SCI as Exemplified by Authors' Profiles

Проблеми ідентифікації метаданих в наукометричних базах даних Web of Knowledge, Scopus і РІНЦ на прикладі профілів авторів

Н. А. Мазов
Институт нефтегазовой геологии и геофизики
им. академика А. А. Трофимука СО РАН, Новосибирск, Россия

В. Н. Гуреев
Государственный научный центр вирусологии и биотехнологии «Вектор»,
Кольцово, Россия

Nikolay Mazov
A. A. Trofimuk Institute of Oil and Gas Geology and Geophysics,
Siberian Division of the Russian Academy of Sciences, Novosibirsk, Russia

Vadim Gureev
«Vector» State Research Center of Virology and Biotechnology,
Koltsovo, Russia

Н. А. Мазов
Інститут нафтогазової геології та геофізики
ім. академіка А. А. Трофимука СВ РАН, Новосибірськ, Росія

В. Н. Гурєєв
Державний науковий центр вірусології та біотехнології «Вектор»,
Кольцово, Росія

В реферативных библиографических базах данных для упрощения обмена информацией используются уникальные идентификаторы для каждого информационного источника, позволяющие легко их отыскивать. В настоящее время нет единого стандартизованного принятого способа идентификации журнальных статей, авторов и др., несмотря на то, что в последние годы введено в действие немалое число различных идентификаторов. Проблема идентификации становится особенно актуальной при использовании наукометрических баз данных, в связи с появлением различных веб-сервисов, позволяющих проводить комплексную обработку данных с дальнейшей интеграцией полученных данных. В настоящей работе рассмотрены инстументарии для идентификации авторов, предоставляемые базами данных Web of Knowledge, Scopus и РИНЦ с позиций эффективности, достоверности и оперативности.

To simplify exchange of information and information retrieval, unique identifiers for each information source are used in abstract bibliographic databases. At present, there is no generally accepted uniform standardized method of identification of magazine articles, authors, etc., despite various identifiers have been implemented in the recent years. The problem of identification becomes particularly urgent when scientometric databases are used, due to the emergence of various web services, allowing carrying out complex data processing with further integration of obtained data. The tools for author identification, provided by Web of Knowledge, Scopus and Russian Science Citation Index databases are examined in the paper in the aspects of promptness, reliability and efficiency.

У реферативних бібліографічних базах даних для спрощення обміну інформацією використовуються унікальні ідентифікатори для кожного інформаційного джерела, що дозволяє легко їх знаходити. На сьогодні немає єдиного стандартизованого способу ідентифікації журнальних статей, авторів та ін., незважаючи на те, що в останні роки введено в дію значну кількість різноманітних ідентифікаторів. Проблема ідентифікації стає особливо актуальною при використанні наукометричних баз даних у зв'язку з появою різноманітних веб-сервісів, які дозволяють проводити комплексну обробку даних із подальшою інтеграцією отриманих даних. У цій роботі розглянуто інструментарій для ідентифікації авторів, що надається базами даних Web of Knowledge, Scopus та РІНЦ з позицій ефективності, достовірності та оперативності.

Для упрощения обмена информацией в реферативных библиографических базах данных (БД) принято использовать уникальные идентификаторы для различных информационных источников, позволяющие легко их отыскивать. При этом в настоящее время в мире нет единого стандартизованного принятого способа идентификации журнальных статей, авторов, их мест работы и др., несмотря на то, что в последние годы введены в действие немалое число различных идентификаторов. Так, в последние десятилетия введены в действие такие международные идентификаторы [1-3], как

- идентификатор издательского произведения – PII (**P**ublisher **I**tem **I**dentifier), введен в 1995 г.;
- идентификатор сериального издания и публикации в нем – SICI (**S**erial **I**tem and **C**ontribution **I**dentifier), введен в 1997 г.;
- идентификатор цифрового объекта – DOI (**D**igital **O**bject **I**dentifier), предложенный Ассоциацией американских издателей;
- международный стандартный код произведения – ISWC (**I**nternational **S**tandard **W**ork **C**ode), предложенный Международной федерацией организаций по правам на копиро-вание;
- унифицированное название источника – URN (**U**niform **R**esource **N**ame)
- и др. разработки.

Авторам практически неизвестны работы, где бы обсуждались проблемы идентификации авторов, профилей авторов, организаций.

Профиль автора – это такой идентификатор в базе данных, который аккумулирует информацию о местах работы автора, количестве его публикаций и их цитируемости, годах публикационной активности, области исследований, соавторах, индексе Хирша, списке использованных в работах литературных источников и пр. Автоматическая обработка данных пока не позволяет однозначно идентифицировать авторов, поэтому вместо одного профиля создается несколько, отчего теряется полнота информации. Причин тому может быть несколько: существование работающих в одной области однофамильцев с одинаковыми инициалами, написание фамилии автора с различным количеством инициалов, смена женских фамилий после замужества, различные варианты транслитерации неанглоязычных фамилий в зарубежных БД, а также опечатки [4].

Современные системы идентификации совершенствуются за счет включения в поиск все большего количества параметров. Например, используется общедоступная информация об авторе в сети Интернет [9], анализируются полученные цитирования и самоцитирования [7], а также информация с первой страницы публикации (заглавие, соавторы, название организации, адреса электронной почты, ключевые слова, резюме) [8]. Тем не менее, лучшими разработками являются полуавтоматические, привлекающие к процессу обработки данных интеллектуальные возможности человека [6].

Отдельным направлением в плане повышения точности информации об авторах стало делегирование им прав на редактирование данных о своих публикациях и цитируемости. Авторам не удалось найти публикаций по этой теме за исключением инструкций на официальных сайтах баз данных, а приведенная информация – результат наших собственных наблюдений в процессе работы с авторскими профилями ученых ГНЦ ВБ «Вектор» и ИНГГ СО РАН.

Сравнительный анализ средств редактирования авторских профилей мы проводим со следующих точек зрения:

- эффективности, т. е. соответствия затраченных автором усилий полученному результату;
- достоверности – каким образом проверяются введенные автором данные;
- скорости – сколько времени занимает процесс редактирования профиля.

Первой рассмотрим надстройку **«My ResearcherID»** от Web of Knowledge (WoK). Для работы в ней автору необходимо зарегистрироваться, регистрация проходит в реальном времени. После этого автор может добавлять статьи в список своих публикаций одним из трех способов:

1) путем поиска в БД WoK и сохранения результатов;
2) посредством загрузки из библиографической программы EndNote;
3) импорт внешних файлов в формате RIS, содержащих статьи, отсутствующие в WoK.

Все эти операции также проходят в реальном времени. Следует отметить, что если автор использует для поиска все базы данных, содержащиеся в WoK, то неизбежно дублирование (так например, одна и та же статья может быть отображена и в Biosis, и в Web of Science, и в Medline, в которых она представлена по-разному). Важно то, что организация списка публикаций в **«My ResearcherID»** не влияет на информацию об авторе в WoK, и проблема множественных авторских профилей не решается. Автор может лишь дать ссылку на уточненный им список своих работ в **«My ResearcherID»**, которая появится под первым авторским множеством. Таким образом, автор может быстро организовать список своих публикаций тремя удобными способами, но эти данные никак не используются разработчиками БД. Thomson Reuters не отразит его в WoK и не станет проверять достоверность предоставленной автором информации. (В ближайшее время Thomson Reuters обещает использовать предоставленную авторами информацию о своих публикациях для уточнения информации в WoK [5]). Из сказанного также следует, что анализ цитирования составленного автором списка будет проводиться только на основе статей, проиндексированных в Web of Science.

В отличие от WoK, БД Scopus и РИНЦ, не имеет специализированной надстройки для авторов. Тем не менее, авторам предоставлены несколько способов уточнения своей библиографической информации. Во-первых, это объединение авторских профилей (request to merge authors), причем запрос на объединение можно сделать как из общего списка авторов, так и из конкретного профиля, где также выводится список профилей, потенциально принадлежащих данному автору. Во-вторых, автор имеет возможность уточнять информацию на уровне статей (request author detail corrections), отмечая свои и исключая чужие публикации во время работы со списком. В окне комментариев можно уточнить организацию, выбрать из предложенных или прописать самому желаемое отображение своего имени. И в-третьих, автор может не пользоваться встроенным инструментарием, а просто выделить в БД Scopus свои статьи и послать текстовый запрос по электронной почте. Заявленные изменения проходят процесс ручной проверки в течение 3 недель.

Таким образом, в отличие от WoK, автор не имеет прямого доступа к редактированию своего профиля, а само редактирование занимает намного больше времени. Однако разработано несколько удобных способов для обратной связи, все изменения отражаются непосредственно в БД, и, что важно, вся введенная автором информация проходит ручную проверку.

У БД РИНЦ перед двумя другими преимущество в отсутствии авторских множеств. Обратной стороной этого является то, что до двух третей работ оказываются «непривязанными» к авторскому профилю, и увидеть их можно, только зайдя в раздел «непривязанных» публикаций (это же касается и цитирований). Как и в WoK, автору предлагается пройти процесс регистрации для получения возможности редактировать свой профиль. Предлагаемая анкета содержит намного больше пунктов, после заполнения которых, анкета проходит модераторскую проверку, на что может уйти до нескольких месяцев (в последнее время наблюдается значительный положительный сдвиг в этой ситуации, и получение SPIN-кода занимает одну-две недели). Отметим, что подача заявки предполагает, кроме проверки анкеты, ручную проверку публикаций автора, так что к моменту получения SPIN-кода у автора будет хорошо отредактированный профиль и у него на данном этапе не будет необходимости вносить какие-либо изменения. Автор может корректировать список своих публикаций и цитирований впоследствии, поскольку в БД добавляются как новые, так и старые публикации. Отметим, что в отличие от **«My ResearcherID»** WoK, работа возможна только с публикациями, проиндексированными в РИНЦ (или из списков литературы из статей БД РИНЦ). Следует указать также на то, что после выдачи автору SPIN-кода модераторская проверка больше не проводится, так что недобросовестный автор теоретически может включить в список своих работ и цитирований публикации и цитирования однофамильцев. А с учетом того, что изменения отражаются в самой БД, указанная возможность впоследствии может негативно сказаться на достоверности представленной в РИНЦ информации.

Подведем итоги нашего обзора. Надстройка WoK **«My ResearcherID»** дает автору наиболее широкие возможности по работе со своими публикациями, позволяя ему использовать ряд различных методов для составления списка своих публикаций, задействовать программное обеспечение для работы с библиографией, управлять доступом к своей странице. Отрицательными моментами выступают отсутствие проверки подлинности заявленной автором информации и относительно

независимое существование надстройки и баз данных WoK. Таким образом, автор не может повлиять на отражение данных о своих публикациях в самих базах данных, а может лишь разместить ссылку на свою страницу. Надстройка Science Index от РИНЦ обладает тем преимуществом, что автору требуется лишь подать заявку на предоставление ему доступа к правке своего профиля, после чего работники технической службы сами отредактирует его профиль. Однако это требует существенных временны́х затрат, ввиду нерасторопности технических служб РИНЦ, а последующие действия автора уже не модерируются, и ответственность за достоверность данных возлагается исключительно на автора.

На наш взгляд, наиболее сбалансированный подход представлен в БД Scopus, в которой авторам напрямую не предоставлены права на внесение изменений, но в которой все продумано для того, чтобы автору легко было сообщить о неточностях и внести нужные коррективы. Наше мнение по поводу наиболее удобного подхода редактирования авторского профиля в БД Scopus подтверждается и авторами [6]. Существенно то, что внесенные автором изменения проверяются и впоследствии отражаются в самой БД, а не в отдельной надстройке, что делает эту БД более точной и удобной в плане поиска информации. Итогом совместной корректировки автором и технической службой БД списка публикаций становится выверенный авторский профиль, который наиболее полно отражает информацию о публикационной активности автора, его цитируемости, сфере деятельности, всех возможных местах работы и пр.

Список литературы:

1. Paskin N. Information identifiers // Learn. Publ. - 1997. - Vol. 10. - № 2. - P. 135.
2. Lupovici C. Identification des ressources sur Internet et metadonnees. Diversite des standards // Doc. Sci. inf. - 1999. - Vol. 36. - № 6. - P. 321-325, 363-364.
3. Бездушный А.Н., Кулагин М.В., Серебряков В.А. и др. Предложения по наборам метаданных для научных информационных ресурсов // Вычислительные технологии - 2005. - Т. 10. - С. 29-48.
4. Egg L., Rousseau R. Introduction to informetrics: Quantitative methods in library, documentation and information science. – Amsterdam: Elsevier science publishers, 1990. - P. 217-218.
5. http://wokinfo.com/5DAF/
6. 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.
7. McRae-Spencer D.M., Shadbolt N.R. Also by the same author: AKTiveAuthor, a citation graph approach to name disambiguation. In Proceedings of ACM/IEEE joint conference on digital libraries (JCDL). 2006, June 11-15, NC, USA. - P. 53-54.
8. Song Y., Huang J., Councill I., et al. Efficient topic-based unsupervised name disambiguation. In Proceedings of the ACM IEEE joint conference on digital libraries (JCDL), 2007, June 18-23, Vancouver, Canada. - P. 342-351.
9. Yao L., Tang J., Li J. A unified approach to researcher profiling // Proceedings of the ieee/wic/acm international conference on web intelligence. - 2007. - P. 359-365.