УДК 004.65:001.89–047.44

И.В. Селиванова, Д.В. Косяков, А.Е. Гуськов

Влияние ошибок в базе данных *Scopus* на оценку результативности научных исследований*

На основе случайной выборки профилей 400 российских авторов и 400 организаций рассматриваются причины возникновения профилей-дублей в базе данных Scopus. Оценивается количество профилей-дублей, анализируется погрешность, которую могут вносить ошибки в библиографических описаниях в результаты наукометрических исследований, основанных на базе данных Scopus. Анализ показал, что в Scopus 76% организаций и 24% авторов имеют профили-дубли. В связи с этим организации теряют в среднем 17% публикаций, авторы – 11%. Результаты исследования могут быть использованы при корректировке базы данных Scopus и оценке погрешности при исследовании результативности научной деятельности.

Ключевые слова: *библиографические базы данных, Scopus, идентификация, наукометрия, библиометрия, библиографические ошибки, ORCID*

ВВЕДЕНИЕ

Среди целей научно-технического развития России, определенных в майских указах Президента РФ в 2012 г., было обозначено увеличение числа российских публикаций в базе данных *Web of Science* (*WoS*) до 2,44% к 2015 г., а также вхождение не менее пяти университетов России в первую сотню ведущих мировых университетов к 2020 г. В связи с этим национальная научная политика все больше ориентируется на количественные показатели публикационной активности российских ученых.

Источником информации для оценки публикационной активности служат библиографические базы данных (ББД), такие как *WoS*, *Scopus*, РИНЦ. Руководство научных организаций использует их при оценке научной результативности сотрудников и при расчете стимулирующих выплат, что мотивирует научных работников публиковать результаты исследований в журналах, индексируемых в международных базах данных.

Национальная и ведомственная оценки результативности научной деятельности организаций, выполняющих научно-исследовательские, опытно-конструкторские и технологические работы также основываются на сведениях о публикационной активности, взятых из этих баз данных [1, 2]. Наукометрические показатели, получаемые из ББД, рассматриваются как истина в последней инстанции, но при работе с ними не учитываются ошибки, влияющие на эти оценки.

Цель нашей статьи – анализ погрешности, которую могут вносить библиографические ошибки в российские исследования, связанные с оценкой результативности научной деятельности, источником данных для которых служит *Scopus*.

ОБЗОР ИССЛЕДОВАНИЙ ПО ПРОБЛЕМЕ

Качество ББД изучалось в некоторых зарубежных работах. Так в статьях F. Franceschini, D. Maisano, L. Mastrogiacomo, посвященных выявлению основных ошибок, встречающихся в *Web of Science* и *Scopus*, выделено два типа подобных ошибок [3, 4].

1. Ошибки, сделанные авторами / издателями / редакторами при подготовке списка литературы:
   - пропущенный или неверный заголовок статьи в списке литературы;
   - ошибки в других полях, таких как имя автора, название журнала, год, том и номер журнала.
2. Ошибки, возникшие в БД при связывании (т.е. установлении идентифицирующих связей между объектами в одной или разных системах, например, публикация – автор, автор – аффилиация) статей:
   - ошибка в имени автора, допущенная при переносе в БД (например, *Özel* переносится как *Oezel*);
   - расширенный список литературы (например, вместо 24-х реальных ссылок в *Scopus* указано 192, из которых 168 возникли в результате «фантомного цитирования», т. е. добавления публикаций, не упоминавшихся в оригинальной версии статьи);
   - полупустой список литературы (вместо библиографического описания указано *Reference information not available*);
   - полностью отсутствующий список литературы;
   - неверный или отсутствующий *Digital Object Identifier* (DOI);
   - потеря цитирования из-за статей вида *Online-first*;

* Работа выполнена в рамках темы научно-исследовательских работ №0334-2019-006 при поддержке Российского фонда фундаментальных исследований, грант № 18-011-00797.

- непроиндексированные статьи (иногда БД «забывают» проиндексировать некоторые статьи, хотя другие публикации из этого же выпуска журнала могут быть проиндексированы. Расчет частотности этой ошибки показал, что она характерна в большей степени для *Scopus*, чем для *WoS*);
- статья проиндексирована в БД, но по неизвестной причине отсутствует в списке литературы процитировавшей ее статьи.

Эту же классификацию ошибок применяет R.A. Buchanan в работе с использованием *Science Citation Index Expanded* и *SciFinder Scholar* [5]. Проблемы со списком литературы обсуждают в своей статье N.J. van Eck и L. Waltman [6].

Исследования показывают, что такие ошибки носят систематический характер и приводят к трудностям при поиске публикаций, а также к сильному искажению библиометрических показателей, относящихся к журналам, ученым или научным организациям. Массовое использование глобальных идентификаторов публикаций *DOI* частично решило проблему идентификации публикаций в пристатейных списках литературы при формировании индексов цитирования. Однако ситуация по-прежнему далека от идеальной: один и тот же идентификатор может встречаться у двух разных статей, одна статья может иметь два разных идентификатора, а сам идентификатор может быть указан неверно (в частности, из-за путаницы символов “O” и “0”, “O” и “Q”, “b” и “6”) [7, 8].

Рассматриваются и ошибки, связанные с появлением дублей статей, которые возникли из-за неверного отнесения статьи к журналам одного и того же издательства, а также из-за орфографических различий и изменений названия журнала [9].

В нескольких публикациях устанавливаются причины возникновения ошибок в именах зарубежных авторов. В работе C. Demetrescu и соавторов [10] разбираются ошибки в написании имен итальянских авторов в четырех базах данных: *WoS*, *Scopus*, *PubMed*, *CrossRef*; в табл. 1 представлены выделенные типы некорректного написания.

S. Ainsworth и J.M. Russell сравнивают соотношение ошибочно написанных имен испанских авторов в различных БД на примерах из латиноамериканского журнала «*Investigación Bibliotecológica*» за 2012 и 2015 гг. [11]. Ошибки в основном касаются порядка отцовских и материнских фамилий, что характерно для испанских авторов. В работе V. Aman на примере лауреатов премии Лейбница оценивается количество дубликатов *Scopus AuthorID* [12].

Общее влияние качества массива данных на проведение библиометрического анализа при составлении авторских рейтингов и регрессионном анализе, где в качестве зависимых переменных выступает число публикаций, оценивает J. Schulz [13]. Результаты этого исследования показали, что ранжирование авторов можно проводить только при очень высоком качестве данных. Хотя рассматривалось только несоответствие имен авторов, J. Schulz приходит к выводу, что ошибки в названиях организаций тоже будут оказывать большое влияние при составлении различных рейтингов (например, рейтингов университетов). Для регрессионного же анализа с наборами библиометрических данных должны проверяться систематические различия в ошибках, которые коррелируют с независимыми переменными (например, если использовать в качестве независимой переменной страну) и могут влиять на достоверность результатов регрессии.

Обобщая зарубежные исследования, можно дать следующую классификацию ошибок в библиографическом описании:

1. Список литературы
   a. Неточность в списке цитируемой литературы
   b. Отсутствие списка литературы
   c. Неверное количество источников в списке литературы
2. Статья
   a. Ошибка в названии
   b. Ошибки цитирования статей из-за публикаций online-first
   c. Ошибки в DOI
3. Журнал
   a. Опечатки в названии, годе издания, томе или номере журнала
   b. Изменение названия журнала
   c. Различные варианты написания названия одного и того же журнала
4. Авторы
   a. Составные фамилия и имя
   b. Диакритические знаки
   c. Апострофы
   d. Опечатки в фамилии и имени

Таблица 1

Типы некорректного написания имен итальянских авторов

Тип некорректного написания	Правильное написание	Некорректное написание
Часть составной фамилии переходит в имя и становится инициалом	De Rossi, Giuseppe	Rossi, G.D.
Часть составного имени становится фамилией	Verdi, Carlo Maria	Maria Verdi, C.
Одна или более частей исчезают из составной фамилии	La Torre, V	Torre, V
Из фамилии исчезают диакритические знаки	Trifrò, S.	Trifro, S.
Диакритические знаки из фамилии отображаются некорректно	Spanò, A.	Span□, A.
Из фамилии исчезает апостроф	D’Innocenzo F.	Dinnocenzo F.
Опечатки в фамилии	Accornero, F.	Accomero, F.
Опечатки в имени	Bianchi, Erica	Bianchi, Enrica

Если ошибки, связанные с журналами, статьями и списками литературы, подходят и к российскому случаю, то случаи 4.a, 4.b и 4.c применимы только к фамилиям зарубежных авторов.

У написания имен, отчеств и фамилий российских авторов и написания названий научных организаций России существует своя специфика. Она связана как с транслитерацией букв кириллического алфавита, не имеющих аналогов в латинице (например, *ё, ю, ш, щ, ъ, ы, ь*), так и с иерархией подчиненности научных организаций в России (имеется в виду принадлежность организаций к государственным академиям наук РФ, либо к их региональным отделениям), а также с ее изменениями в последние годы. Эти ошибки служат причиной неверной привязки публикаций, что в свою очередь ведет к появлению профилей-дублей авторов и организаций.

С целью однозначного определения фамилий авторов в работе [14] предложена система уникальных идентификаторов *ORCID*, в последние годы широко признанная научным сообществом и принятая многими издателями научных журналов. Этот идентификатор не завязан на владельце ББД, используется как в *Scopus*, так и в *WoS* и *Dimensions*. Таким образом, *ORCID* может быть идентификатором, претендующим на глобальное признание [15].

Многие журналы требуют указание идентификатора *ORCID* в списке авторов научной статьи. *ORCID* используется и в национальных системах научной информации [16]. Ранее компанией *Thomson Reuters* был предложен универсальный идентификатор автора *ResearcherID*, который не получил широкого признания в связи с тесной связкой с базой данных *WoS*. Компания *Elsevier* в системе *Scopus* использует уникальные идентификаторы авторов *Scopus AuthorID*. Все эти идентификаторы часто применяются в различных информационно-аналитических системах научных институтов при интеграции данных о публикационной активности их сотрудников [17]. Но при связывании данных все равно отмечается необходимость их визуальной проверки [18].

К сожалению, сами авторы плохо отслеживают свои публикации [19], что приводит к неполноте их профилей в ББД, и, соответственно, к некорректным результатам связывания при использовании систем идентификации.

Предпринимаются попытки и для введения уникальных идентификаторов организаций. Одним из таких проектов является *Research Organization Registry* (ror.org), инициированный *California Digital Library*, *Crossref*, *DataCite* и *ORCID*. Однако этот проект находится в ранней стадии, а предыдущие инициативы, такие как идентификатор организации *ISNI* [20], не были широко приняты сообществом издателей и владельцев ББД.

МЕТОДЫ ИССЛЕДОВАНИЯ

Данные

Для нашего анализа была выбрана БД *Scopus*, в которой у каждого автора и организации может быть свой профиль с идентификатором. Данные о публикациях российских авторов были получены с помощью *Scopus Search API* и *Scopus Abstract Retrieval API* в 2017 г.

Дубли организаций

Наличие профиля-дубля у российских организаций проверялось на случайной выборке из 400 профилей. Это количество было получено из формулы В.И. Паниотто с тем условием, что доверительная вероятность P=0,954. Для указанной доверительной вероятности коэффициент доверия $t = 2$, дисперсия качественного альтернативного признака принимается максимально возможной, т.е. соответствующей доле 0,5 [21, с.62]:

$$n = \frac{t^2 p(1-p)}{\Delta^2} = \frac{2^2 \cdot 0{,}5(1-0{,}5)}{\Delta^2} = \frac{1}{\Delta^2} = \frac{1}{0{,}05^2} = 400 , \quad (1)$$

где n – объем выборки; p – генеральная доля единиц, обладающих значением признака, относительно которого рассчитывается ошибка выборочной доли; t – доверительный коэффициент; Δ – предельная ошибка выборки.

Для того чтобы анализ был более точным, введено ограничение на количество публикаций в профиле организации: в 2017 г. их должно быть не менее 5. При меньшем количестве публикаций в список попадали организации, реальное название которых сложно было определить. Например, организация с *Affiliation ID*=119911371 называлась *Berdsk*, получившееся из-за перехода в название части адреса. Также были обнаружены явные дубли других организаций. В качестве примера приведем организацию с *Affiliation ID*= 109886056, в названии которой указано *Romsk Polytechnical University*. Можно предположить, что этот профиль, скорее всего, является дублем профиля Томского политехнического университета (ТПУ). Подтвердить это предположение невозможно, так как в этом профиле только одна публикация, в аффилиациях авторов которой указана только эта организация. В основном профиле ТПУ этих авторов обнаружить не удалось.

Алгоритм анализа профилей организаций

1. Для организации формируем список из 10 аффилированных с ней авторов с наибольшим числом публикаций: $A_i = \{a_1, a_2, \ldots, a_M\}$, где M = 10.
2. Для каждого автора a_i, i=1,2, …, M формируем полный список его аффилиаций $OA_i = \{oa_{i1}, oa_{i2}, \ldots\}$.
3. Формируем общий список аффилиаций 10 самых продуктивных авторов из организации $GL = \bigcup OAi$.
4. В списке *GL* выявляем дубликаты – переводы названий, сокращенные и схожие названия.

Алгоритм применялся для каждой из 400 организаций и позволил выявить дублирующие профили аффилиаций. При сопоставлении названий дополнительно изучались сайт организации, «Википедия», а также профиль организации в Научной электронной библиотеке eLibrary.ru.

Дубли авторов

Наличие профиля-дубля у автора проверялось по формуле (1) на случайной выборке из 400 активных российских авторов (не менее двух публикаций в 2017 г.). Ограничение на количество публикаций бы-

ло введено из-за того, что при отсутствии этого фильтра в случайную выборку попадало слишком много профилей авторов с одной публикацией, что не позволяло проанализировать наличие у этих авторов дублирующих профилей. В списках российских авторов из Scopus было обнаружено около 47% таких профилей.

Алгоритм поиска профиля-дубля автора

Обозначим множество авторов за $A = \{a_1, a_2, \ldots, a_N\}$, где $N = 400$.

1. Для каждого автора a_i, i=1, …, N формируем полный список аффилиаций $OA_i = \{oa_{i1}, oa_{i2}, \ldots\}$, т. е. для каждого из авторов – полный список организаций, о которых есть упоминание в его публикациях.
2. Для каждой организации oa_{ij} формируем полный список авторов $C_{ij} = \{c_{ij}^1, c_{ij}^2, \ldots\}$, т. е. авторов, работающих в одной организации o_{ij} с автором a_i, включая и самого автора a_i.
3. Формируем общий список коллег автора a_i из всех организаций $C_i = \bigcup_j C_{ij}$.
4. Из списка C_i находим все дубликаты – авторы с ФИО автора a_i, которое либо полностью совпадает с его ФИО, либо похоже на него. Анализ проводился вручную, и совпадения проверялись экспертным путем (при автоматизированном поиске необходимо использовать различные меры близости).

Для того чтобы исключить возможность определения профиля полного тезки автора как профиля-дубля исследуемого автора, дополнительно анализировались списки соавторов, тематики исследования, организаций, полные тексты статей, списки авторов на eLibrary.ru и аффилиационная история. Мы полагаем, что если у нескольких профилей с одинаковыми ФИО совпадают места работы, соавторы и тематика, то это профили одного и того же автора.

Этапы исследования

Каждый из трех этапов исследования выполнялся как для российских авторов, так и для организаций.

1. Определение причины возникновения профилей-дублей
2. Оценка количества таких профилей
3. Оценка доли публикаций в профилях-дублях от общего количества.

На каждом этапе результаты были получены и для организаций, и для авторов.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Оценка количества и определение причин возникновения профилей-дублей организаций

Анализ профилей организаций показал следующие результаты.

A. 19% (75 из 400) организаций дублей не имеет.

B. 5% (20 из 400) профилей организаций содержат ошибки (иностранные организации, разные адреса и др.), что затрудняет определение дублей.

C. 76% (305 из 400) организаций имеют хотя бы один дубль.

Рассмотрим подробнее ошибки, обнаруженные в профилях 5% организаций (результат ***B***).

Случай B1. *Вместо названия организации указаны профили: РАН, СО РАН и др.*

Ошибка возникает из-за неверной привязки организации, в названии которой встречается, например, Сибирское отделение Российской академии наук (СО РАН). Такая организация может быть отнесена к профилю как СО РАН, так и РАН, при этом в последнем случае публикация будет приписана Москве, а не Новосибирску или другому сибирскому научному центру. Это может повлиять на результаты некоторых исследований, например, по академической мобильности ученых (*Affiliation ID* профиля СО РАН в *Scopus* – 60017604, РАН – 60021331).

Для оценки количества публикаций, которые может потерять организация из-за отнесения к профилям РАН, РАМН или их региональным отделениям, была сделана случайная выборка 400 профилей российских авторов и посчитана доля публикаций для случая B1. Результаты показали, что из-за неверного отнесения публикаций авторов к государственным академиям наук РФ в среднем для одного автора организация теряет до 14% публикаций.

Случаи, когда автор действительно может указывать аффилиацию Российской академии наук или ее региональных отделений, являются редкими.

Случай B2. *Иностранные организации*

Ошибка возникает из-за неверного распознавания названия города. Например, университет Айдахо расположен в городе *Moscow*, штат Айдахо, США. Но в *Scopus* один из профилей университета относится к России (*Affiliation ID*=100804052, количество публикаций = 106). При этом в список из 400 организаций попал еще один профиль университета Айдахо, также отнесенный к российским организациям (*Affiliation ID*=110070192, количество публикаций = 73).

Случай B3. *Сложносоставные организации с разными подразделениями*

К данному случаю относятся как университеты, где авторы иногда указывают только название факультета, и такая публикация может попасть в другой профиль, так и организации, подразделения которых находятся в разных городах. В качестве примера приведем организацию с *Affiliation ID*=100465640. В ее названии указано *Chemistry Department*. При этом в адрес организации попало полное название *Lomonosov MSU*, *Moscow*.

В табл. 2 указаны причины возникновения дублей профилей российских организаций (результат C), а также частота их встречаемости на примере тех профилей, которые имеют только один дубль (общим количеством 66).

Данные в ББД *Scopus* активно обновляются и их качество растет, как в связи с усилиями специалистов *Elsevier*, так и с обращениями пользователей системы. Результаты нашего исследования показали, что из 66 дублирующихся профилей на начало 2017 г. почти 41% оказались объединенными к марту 2019 г.

Таблица 2

Анализ профилей с одним дублем

Случай	Причина возникновения дублей	Кол-во профилей	Доля в общем количестве, %
C1	Старое название организации	1	2
C2	Сокращенное название организации, включая • частичную аббревиацию названия (например, вместо *the Siberian Branch of the RAS* указано *SB RAS*); • сокращенное название, добавленное к полному	13	20
C3	Отсутствие части названия	17	26
C4	Отсутствие имени/части имени ученого, присвоенного организации, его инициалов	3	5
C5	Попадание части названия организации в адрес	2	3
C6	Неверный/неполный адрес организации (включая неверное указание страны или города)	4	6
C7	Разный порядок слов в названии	1	2
C8	Полное совпадение названий (например, у *Tver State Medical University* имелся дубль с таким же названием)	4	6
C9	Разная транслитерация названия	2	3
C10	Разный перевод названия (включая перевод не только на английский язык)	9	14
C11	Ошибки в названии (опечатки, лишние пробелы и знаки препинания)	10	15
C12	Зависимость от регистра	0	0

Таблица 3

Профили-дубли Национального медицинского исследовательского центра психиатрии и наркологии имени В.П. Сербского

Affiliation ID	Название организации	Ошибка
60085185	Serbsky Institute for General and Forensic Psychiatry	C1
100312210	Serbsky Res. Inst. Forens. Psychiat.	C1, C2
100331806	V. P. Serbskii Central Research Institute of Forensic Psychiatry	C4, C9, C11
100360057	Serbsky National Research Center for Social and Forensic Psychiatry	C1
100366550	Serbsky Natl. Res. Ctr. Social F.	C1, C2
100368625	V. P. Serbskii State Research Center	C9, C10
100390434	Gos. Nauchnyj Tsentr Sotsial'noj	C3
100755591	NI Inst. Obshchej/Sudebnoj Psikhiat.	C1, C2, C4, C9
101328961	Serbskii State Scientific Center for Social and Forensic Psychiatry	C1, C9
101335050	Lab. of Fed. State Inst. State Sci. Centre of Social and Forensic Psychiatry after V.P. Serbsky	C1, C7
101981345	V. P. Serbskii State Research Center of Social and Forensic Psychiatry	C1, C4, C9
105353230	Serbsky National Research Centre for Social and Forensic Psychiatry	C1
108509206	Serbsky's Institute of General and Forensic Psychiatry	C1, C10
108847913	Serbsky National Research Centre for Social and Forensic Psychiatry	C1
112568830	V. P. Serbskii State Scientific Center for Social and Forensic Psychiatry	C1, C4, C9, C11
112615739	Serbsky State Research Center of Social and Forensic Psychiatry	C1
112617371	V. P. Serbsky State Research Center for Social and Forensic Psychiatry	C1, C11
112625607	V. P. Serbskii Center of Social and Forensic Psychiatry	C1, C4, C9, C11
112662194*	V. P. Serbskii State Research Center of Forensic	C1, C3, C4
112956380	V.P. Serbsky State Scientific Center for Social and Forensic Psychiatry	C1, C4
113068622	Serbsky National Research Center for Social and Forensic Psychiatry	C1
113157802	Serbsky State Scientific Center for Social and Forensic Psychiatry	C1
113939370	Serbsky National Research Center	C3
114190516	Serbsky State Scientific Center of Social and Forensic Psychiatry	C1
115403520	Serbsky Federal Medical Research Center of Psychiatry and Narcology	C10
116873944	Serbsky Federal Medical Research Center for Psychiatry and Narcology	C10

* профиль, отсутствующий в *Scopus* на 04.05.19

Максимальное количество – 26 дублей обнаружено в *Scopus* в названии Федерального государственного бюджетного учреждения «Национальный медицинский исследовательский центр психиатрии и наркологии имени В.П. Сербского» Министерства здравоохранения Российской Федерации, сокращённо – ФГБУ «НМИЦПН им. В.П. Сербского» Минздрава России, старые названия которого: «Государственный научный центр социальной и судебной психиатрии им. В.П. Сербского», «Федеральный медицинский исследовательский центр психиатрии и наркологии имени В. П. Сербского. Все найденные профили этой организации представлены в табл. 3. Основным считаем профиль, имеющий максимальное число публикаций.

Оценка количества и определение причин возникновения профилей-дублей авторов

Анализ профилей авторов показал следующие результаты.

A. У 75% (300 из 400) авторов не было обнаружено профилей-дублей.

B. У 1% (3 из 400) не удалось установить наличие дубля.

Это связано, например, с тем, что в профиле организации в *Scopus* встречается несколько авторов с одним и теми же фамилией и инициалами, но в профиле организации, как и на сайте института, такого сотрудника нет. В профилях этих авторов не совпадают также и соавторы. В другом случае имя и фамилия автора имеют вид Alexander R., т. е. определить фамилию этого автора становится невозможным.

C. У 24% (97 из 400) были обнаружены профили-дубли.

Необходимо отметить, что с начала 2017 г. к марту 2019 г. 9% (36 из 400) профилей авторов объединены в *Scopus* полностью, 2% (7 из 400) – частично, а 13,5% (54 из 400) профилей так и остаются необъединенными.

Рассмотрим причины и частоту возникновения 61 профиля-дубля из тех, которые полностью (54 профиля) или частично (7 профилей) не объединены в *Scopus*.

Случай 1. *Разная транслитерация (41 из 61, 67%)*

Разная транслитерация фамилии, имени или отчества автора появляется из-за наличия в них букв *ё, ю, я, х, ц, ш, ь* и др. (например, имя одного автора транслитерируется в одном случае как Yury, в другом – Yrii).

Примером автора, в дублирующих профилях которого наблюдается разная транслитерация и в имени, и в фамилии, является Bakaeva Natalia Vladimirovna (*AuthorID* = 57195920506). В ее профиле-дубле (*AuthorID* = 56826095700) имя указано как Bakayeva, Natalya. Для этого примера отметим еще один важный момент: обоим профилям дается один и тот же *ORCID*, но профили не объединены, т. е. при объединении профилей в *Scopus* механизм объединения по совпадающему *ORCID* либо отсутствует, либо срабатывает не всегда.

Случай 2. *Ошибка в фамилии (10 из 61, 16%).*

- ***Лишние буквы, опечатки, ошибки распознавания, похожие по внешнему виду буквы (например, заглавная I и строчная l), строчные и заглавные буквы, ошибки в самой публикации.***

Одной из иллюстраций этого типа ошибок является профиль автора (основной *AuthorID* = 7202813119) Костромского государственного университета – Smirnova, N.A. Второй ее профиль (*AuthorID* = 56568452600) обозначен как Smimova, N. A., т.е. буквы rn в фамилии были распознаны как m. В ее третьем профиле (*AuthorID* = 55480627000) в качестве ФИО указано: SMlRNOVA, N. A., т.е. буква I была распознана как l.

В другом примере ошибочное написание фамилии автора возникло уже в исходной публикации. Среди профилей авторов Научно-исследовательского института фармакологии им. В.В. Закусова были обнаружены два автора: Seredenin, Sergey B. (*AuthorID* = 7004680975) и Seredin, S. B. (*AuthorID* = 6504406983). При этом в списке авторов на сайте eLibrary.ru автора с фамилией Середин обнаружено не было. В полном тексте статьи из профиля автора Seredin, S. B. также была указана фамилия Середин. Через eLibrary.ru было выяснено, что статья является переводной версией русскоязычной публикации, в которой автором указан Середенин С.Б. О том, что эта статья принадлежит именно Середенину, также свидетельствует наличие общих соавторов с основным профилем, список литературы, где практические все работы принадлежат Середенину.

- ***К окончанию фамилии приписана буква от аффилиации.***

Примером может послужить профиль автора из Института общей и неорганической химии имени Н.С. Курнакова РАН – Fedorchenko, Irina Valentinovna (*AuthorID* = 24474356600). У нее есть профиль-дубль с одной публикацией, где в имени профиля написано Fedorchenkob, I.V. Детальный анализ полного текста этой публикации показал, что *b* в конце фамилии появилась из-за обозначения этой буквой ссылки на организацию автора; в полном тексте публикации данной ошибки нет.

Случай 3. *Разный вариант представления имени (8 из 61, 13%).*

- ***Вместо полного имени/отчества указаны инициалы; отчество отсутствует полностью.***

Рассмотрим профиль автора из Института общей физики имени А.М. Прохорова РАН Kozlov, D.N. (*AuthorID* = 55407907300). В профиле у автора 75 публикаций. Для этого ученого был найден еще один профиль: Kozlov, Dimitry N. (*AuthorID* = 26650432300) с одной публикацией. Соавторы и тематики исследований в профилях совпадают.

- ***Перепутаны местами имя, отчество и фамилия.***

У автора из Университета ИТМО – Leonov Mikhail Yu – два профиля: в первом (*AuthorID* = 36604774600) в качестве фамилии указано: Leonov, имени: Mikhail Yu; во втором (*AuthorID* = 57192376476) фамилия у автора уже Leonov Mikhail, в имени указаны лишь инициал отчества Yu. В основном профиле у автора 39 публикаций, в дубле – одна.

Другим подобным примером является профиль автора из Института гидродинамики имени М.А. Лаврентьева СО РАН, в основном профиле которого (*AuthorID* = 6603685354) указаны фамилия: Liapidevskii, имя: Valery Yu. В одном из его профилей-дублей (*AuthorID* = 57196437043) в фамилию также вошел инициал от отчества: Yu Liapidevskii, а имя осталось верным: V.

Случай 4. *Разные периоды публикаций (1 из 61, 2%).*

Единственным примером, относящимся к этому случаю, является профиль автора из Рязанского государственного медицинского университета. В одном профиле (*AuthorID* = 6505788719) охват публикаций с 1985 по 1993 гг., в другом (*AuthorID* = 57195313658) – с 2017 по 2018 гг. Профиль автора был найден также в eLibrary.ru, где охват публикаций с 1986 по 2018 гг. С профилем в eLibrary.ru совпадают списки статей и большинство соавторов как из первого, так и из второго профиля. Стоит отметить, что на eLibrary.ru возможны, хотя и редки, случаи неверного объединения профилей авторов, но в этом случае из-за совпадения большинства соавторов профиль объединен верно. Дополнительного третьего профиля в *Scopus*, в котором могли бы находиться публикации автора за неотраженный период (с 1994 по 2016 гг.), обнаружить не удалось.

Оценка доли публикаций в профилях-дублях от общего количества

На третьем этапе исследования для оценки количества публикаций, которые могут пропасть из основного профиля, был введен показатель $L = \frac{S(D)}{S(O+D)}$, где $S(D)$ – суммарное число публикаций во всех профилях-дублях, а $S(O+D)$ –общее число публикаций в основном профиле и его профилях-дублях.

Анализ показал, что для организации средние потери публикаций могут составлять 17% (максимальное значение – 83% достигается для уже упомянутого *Serbsky Federal Medical Research Centre for Psychiatry and Narcology*, *Affiliation ID* основного профиля = 116873944), для авторов – 11% (максимальное значение – 55%, *AuthorID* основного профиля = 8630661000).

ЗАКЛЮЧЕНИЕ

Проведенное нами исследование выявило, что 76% организаций и 24% авторов научных публикаций имеют профили-дубли в *Scopus*. Дубли профилей организаций возникают по следующим причинам: различия при транслитерации и переводе названий организаций, опечатки и ошибки в названиях, отсутствие части названия, указание разных адресов, неверное указание принадлежности к тому или иному региональному отделению Российской академии наук.

Дубли профилей авторов также возникают из-за различий при транслитерации и опечаток. Характерной причиной является ошибочное написание фамилии, имени или отчества: к фамилии приписывается символ, взятый из обозначения организации или из отчества; фамилия и имя могут быть перепутаны местами. Были обнаружены и случаи, когда вместо фамилии указано сокращение, что не позволяло идентифицировать автора.

Из-за подобных ошибок в *Scopus* искажается количественная оценка результативности: организации в среднем теряют 17% публикаций, а авторы – 11%.

Наш анализ дает дополнительную информацию для корректировки ошибок в профилях авторов и организаций, которая постоянно происходит в *Scopus*. Вместе с тем, при наукометрическом анализе оценке результативности научной деятельности на основе данных *Scopus* необходимо учитывать погрешность, вызванную некорректной идентификацией и связыванием авторов, организаций и публикаций.

СПИСОК ЛИТЕРАТУРЫ

1. Гуськов А.Е., Косяков Д.В., Селиванова И.В. Методика оценки результативности научных организаций // Вестник Российской академии наук. – 2018. – Т. 88, № 5. – С. 430-443.
2. Kosyakov D., Guskov A. Research assessment and evaluation in Russian fundamental science // Procedia Computer Science. – 2019. – Vol. 146. – P. 11-19
3. Franceschini F., Maisano D., Mastrogiacomo L. Empirical analysis and classification of database errors in Scopus and Web of Science // Journal of Informetrics. – 2016. – Vol. 10, Issue 4. – P. 933-953.
4. Franceschini F., Maisano D., Mastrogiacomo L. The museum of errors/horrors in Scopus // Journal of Informetrics. – 2016. – Vol. 10, Issue 1. – P. 174-182.
5. Buchanan R.A. Accuracy of cited references: The role of citation databases// College & Research Libraries. – 2006. – Vol. 67, Issue 4. – P. 292-303.
6. Nees Jan van Eck, Ludo Waltman. Accuracy of citation data in Web of Science and Scopus // arXiv:1906.07011. – URL: https://arxiv.org/ftp/arxiv/papers/1906/1906.07011.pdf
7. Franceschini F., Maisano D., Mastrogiacomo L. Errors in DOI indexing by bibliometric databases // Scientometrics. – 2015. – Vol. 102, Issue 3. – P. 2181–2186.
8. Zhu J., Hu G., Liu W. DOI errors and possible solutions for Web of Science // Scientometrics. – 2019. – Vol. 118, Issue 2. – P. 709–718.
9. Valderrama-Zurián J.-C., Aguilar-Moya R., Melero-Fuentes D., Aleixandre-Benavent R. A systematic analysis of duplicate records in Scopus// Journal of Informetrics. – 2015. – Vol. 9, Issue 3. – P. 570-576.
10. Demetrescu C., Ribichini A., Schaerf M. Accuracy of author names in bibliographic data sources: an Italian case study // Scientometrics. – 2018. – Vol. 117, Issue 3. – P. 1777–1791.
11. Ainsworth S., Russell J.M. Has hosting on science direct improved the visibility of Latin American scholarly journals? A preliminary analysis of data quality // Scientometrics. – 2018. – Vol. 115, Issue 3. – P. 1463–1484.
12. Aman V. Does the Scopus author ID suffice to track scientific international mobility? A case study

based on Leibniz laureates // Scientometrics. – 2018. – Vol. 117, Issue 2. – P. 705-720.
13. Schulz J. Using Monte Carlo simulations to assess the impact of author name disambiguation quality on different bibliometric analysis // Scientometrics. – 2016. – Vol. 107, Issue 3. – P. 1283–1298.
14. Haak L.L., Fenner M., Paglione L., Pentz E., Ratner H. ORCID: A system to uniquely identify researchers // Learned Publishing. – 2012. – Vol. 25, Issue 4. – P. 259-264.
15. Mazov N.A., Gureyev V.N. Modern challenges in bibliographic metadata identification. 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC), Vladivostok // IEEE. – 2018. – P. 1-4.
16. Moreira J.M., Cunha A., Macedo N. An ORCID based synchronization framework for a national CRIS ecosystem // F1000Research. – 2015. – Vol. 4. – P. 181.
17. Альперин Б.Л., Ведягин А.А., Зибарева И.В. SciAct – информационно-аналитическая система Института катализа СО РАН для мониторинга и стимулирования научной деятельности // Труды ГПНТБ СО РАН. – 2015. – Т. 9. – С. 95-102.
18. Ковязина Е.В. Корпоративные репозитории научных публикаций и проблемы обмена данными // Труды ГПНТБ СО РАН. – 2016. – Т. 10. – С. 288-292.
19. Захарова С.С., Гуреева Ю.А. Научные публикации: от картотеки трудов до библиографических профилей // Библиосфера. – 2017. – №2. – С.85-89
20. MacEwan A., Angjeli A., Gatenby J. The international standard name identifier (ISNI): The evolving future of name authority control // Cataloging and Classification Quarterly. –2013. – Vol. 51, Issue (1-3). – P. 55-71.
21. Могильчак Е.Л. Выборочный метод в эмпирическом социологическом исследовании: учеб. пособие. – Екатеринбург: Изд-во Уральского ун-та, 2015. – 120 с.

Материал поступил в редакцию 05.07.19

Сведения об авторах

СЕЛИВАНОВА Ирина Вячеславовна – младший научный сотрудник ГПНТБ СО РАН
e-mail: selivanova@spsl.nsc.ru;

КОСЯКОВ Денис Викторович – зам. директора по развитию ГПНТБ СО РАН
e-mail: kosyakov@spsl.nsc.ru

ГУСЬКОВ Андрей Евгеньевич – кандидат технических наук, директор ГПНТБ СО РАН
e-mail: guskov@spsl.nsc.ru