



ПОДХОДЫ АВТОМАТИЗАЦИИ ОБРАБОТКИ ДАННЫХ НАУКОМЕТРИЧЕСКИХ БАЗ ДАННЫХ

Вареников Денис Анатольевич, Муромцев Дмитрий Ильич, Шлей Михаил Дмитриевич

Аннотация

В статье представлены результаты разработки подхода для автоматизации процесса загрузки данных из наукометрических баз данных в базу знаний университета. Описаны основные проблемы обработки информации о публикациях. Особое внимание в статье уделено вопросу идентификации авторов публикаций, являющихся сотрудниками университета. Реализация предложенного подхода в информационной системе университета позволила оптимизировать трудозатраты сотрудников по наполнению базы знаний университета, а также повысить качество хранимых данных.

Ключевые слова: наукометрические базы данных, автоматизация, публикация, авторы, соавторы, *Web of Science*, *Scopus*, РИНЦ, идентификация авторов, аффилиация, информационная система.

1. ВВЕДЕНИЕ

Развитие системы проектной деятельности университета предполагает решение ряда задач, одна из которых — формирование базы знаний результатов проектной и научно-исследовательской деятельности сотрудников и подразделений вуза [1, 6]. К основным результатам научно-исследовательской деятельности университета относятся публикации, выступления на конференциях, объекты интеллектуальной собственности (право использования, которых защищено охранными документами). Также к результатам можно отнести информацию о научных исследованиях и разработках, проводимых сотрудниками, руководство научно-исследовательскими работами студентов и диссертационными работами аспирантов.

Сложность решения данной задачи связана с наличием множества источников возникновения данных о результатах научно-исследовательской деятельности, множество информационных систем, куда данные вводятся и где они публикуются. Проектную деятельность ведут студенты, преподаватели, сотрудники, подразделения, в том числе кафедры, научные лаборатории и собственно университет в целом [1, 7]. Наиболее используемыми результатами научно-практической деятельности являются публикации. Ими отчитываются по проектной деятельности в рамках бюджетного и целевого финансирования, НИР и НИОКР [2]. Информация о публикациях преподавателей, которой, прежде всего, отчитываются они сами, может быть включена в результаты проектной деятельности кафедры, научной лаборатории, университета. Ввод и обработка сведений о пуб-

ликациях в больших объемах — довольно трудоемкий процесс. Автоматизация работы по обработке сведений о публикациях, поступающих в информационную систему университета, позволит повысить скорость ее обработки, обеспечить формирование более качественной базы знаний, оптимизировать трудозатраты сотрудников, преподавателей и студентов по формированию различных отчетов.

В данной статье представлены подходы к автоматизации поступления сведений о научно-практических результатах в информационную систему университета. Рассмотрены основные проблемы, возникающие при обработке данных, поступающих из сторонних систем, а также методы их решения.

2. Постановка задачи

Источники данных о публикационной деятельности для информационной системы можно разделить на две категории:

1. Персональный ввод сведений студентами, преподавателями и сотрудниками.
2. Автоматизированная загрузка сведений из внешних авторитетных баз публикаций, в том числе и из наукометрических баз данных. Под наукометрическими базами данных понимают библиографические и реферативные базы данных, а также инструмент для отслеживания цитируемости научных статей [9].

Каждая из категорий возникновения сведений о публикациях обладает своими преимуществами и недостатками. Свободный ввод сведений о публикациях авторами в свое портфолио, зачастую приводит к неполноте введенных данных и их некорректности. Как правило, ввод таких сведений стимулируется необходимостью предоставления различного рода отчетов, проведением конкурсов или материальным стимулированием публикационной деятельности. В связи с ограниченностью времени и большим объемом информации, вводимые сведения в портфолио могут дублироваться, возможен ввод некорректной информации, также возможно указание не всего авторского коллектива, что влечет за собой отсутствие данной публикации у других авторов и, как один из возможных вариантов, внесение дубликата данными авторами. Какие бы защиты, препятствующие вводу дубликатов в систему, ни существовали, пользователи всегда смогут ввести дублирующую по смыслу информацию о публикации. Соответственно, для того чтобы собранные данные о публикациях можно было использовать для построения статистических отчетов, необходимо выполнять специализированную обработку данных [10]. Регулярная автоматизированная загрузка сведений о публикациях из сторонних авторитетных баз данных позволяет значительно упростить работу студентов, преподавателей и сотрудников университета. Данный подход является оптимальным способом наполнения единой университетской базы данных научно-практической деятельности (централизованное хранение). Централизованное хранение — технологические решения для организации взаимодействия между информационной системой университета и внешними информационными системами (например электронными научными изданиями) [3].

3. Основные проблемы идентификации данных

Так же, как и в случае свободного ввода сведений о публикациях, у автоматизированной загрузки данных есть свои трудности. Обычно различные наукометрические

базы данных, такие как E-Library, российская научная электронная библиотека (<http://elibrary.ru>), Web of Science, корпорации Thomson Reuters или Scopus, издательства Elsevier и др., хранят информацию о публикациях и обеспечивают интерфейсы для выгрузки метаданных [8]. Данные из различных источников (внешних баз данных) могут быть организованы и представлены различным образом, что приводит к неопределенностям. Можно выделить следующие неопределённости:

- Структурная неопределенность — сведения в различных базах данных могут быть представлены различными структурами (сведения об авторах, описание издания, представление выходных данных).
- Техническая неопределенность — реализация различных внешних баз данных публикаций организована на различных технологических платформах.
- Неоднородность подходов к предоставлению данных — в зависимости от рассматриваемой внешней базы данных, осуществляется тот или иной доступ к информации и метод ее обработки.

Многие наукометрические базы данных содержат одинаковые издания, к примеру, некоторые издания могут входить в базу данных Web of Science и Scopus, появление сведений о публикациях в этих базах данных может быть различным во времени, что потенциально может породить дубликаты при загрузке данных. Для упрощения обмена информацией в реферативных библиографических базах данных 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 введено в действие немалое число различных идентификаторов [4]. Рассмотрим наиболее распространенные наукометрические базы данных и их идентификаторы авторов:

- РИНЦ (российский индекс научного цитирования) — используется уникальный идентификатор SPIN-код [5].
- Web of Science — самая авторитетная в мире база данных по научному цитированию института научной информации (Institute of Scientific Information — ISI) — используемый уникальный идентификатор ResearcherID [5].
- Scopus — это крупнейшая в мире единая мультидисциплинарная реферативная база данных, представляющая уникальную систему оценки частоты цитирования. Используемый уникальный идентификатор ORCID [5].

Один из возможных подходов идентификации авторских коллективов из различных баз данных публикаций — это проведение анализа возможных внешних идентификаторов авторов и сопоставление их с внутренними (университетскими) идентификаторами. Такие связи идентификаторов не всегда существуют, возникают новые авторские коллективы, автор может изменить фамилию, также в авторитетных базах данных авторы могут быть не привязаны к уникальному идентификатору, или один и тот же автор может быть привязан к разным идентификаторам. При идентификации авторских коллективов большое значение имеет аффилиация авторов. Некоторые авторы не указывают аффилиацию с университетом, что приводит к затруднению их идентификации. В случае работы с аффилиациями можно выделить следующие возможные варианты:

- Указания аффилиации — автор является сотрудником университета и указал ссылку на университет.

- Отсутствие аффилиации — автор является сотрудником университета и не указал ссылку на университет. — автор является сотрудником университета и указал ссылку на несколько университетов.

3.1. Общий подход к автоматизации загрузки данных

Для получения сведений о публикациях сотрудников университета из внешних баз данных был разработан подход, суть которого заключается в том, что обработка данных разбивается на несколько этапов. Обработка данных в рамках каждого этапа выполняется при помощи соответствующего модуля. На рис. 1 представлена общая поэтапная схема обработки данных.

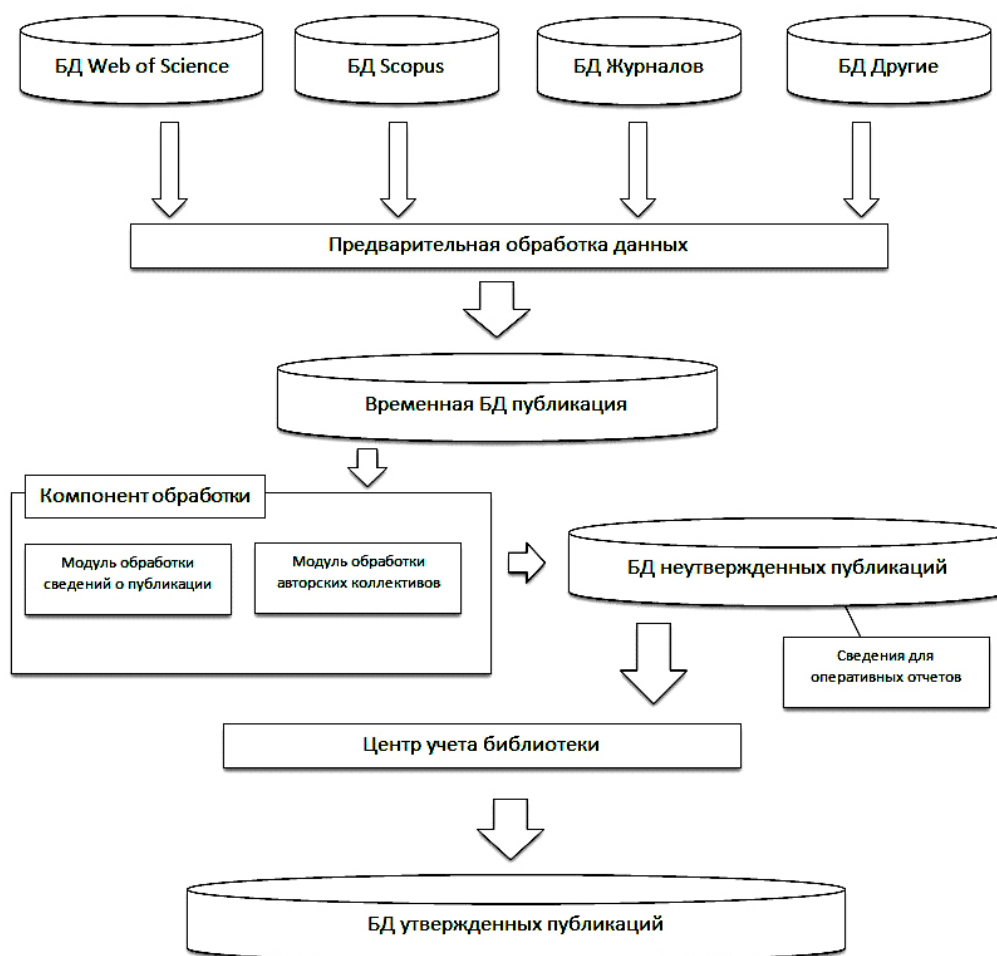


Рис. 1. Подход к автоматизации процесса загрузки данных

На первом этапе выполняется предварительная обработка данных с целью приведения сведений о публикациях к единой структуре. Данный этап необходим в связи с тем, что данные, экспортируемые из внешних источников, имеют различную структуру. Для примера рассмотрим структуру выгружаемой информации о публикации из наукометрической базы данных Scopus и Web of Science. Фрагмент экспортируемого файла из базы данных SCOPUS:

```

<div> <span> Mouromtsev, D., Pshenichny, C., Yakovlev, A. </span> <br>
<a target="_blank" title="go to record page" href="http://www.scopus.com/
record/display.url?eid=2-s2.0-84892371567&origin=resultslist">
<span> Semantic and structural delineation of market scenarios by the event bush
method </span> <br> </a> <span> (2014) &nbsp;<em> Decision Support Systems </em>
, &nbsp;<span> 57 &nbsp;</span> (1), &nbsp;<span> pp. &nbsp;</span> 34-41. &nbsp;<br> </span> </div>
<div> <span> <br> Intellectual Systems Laboratory, National Research University
of Information Technologies, Mechanics and Optics, Kronverksky Prospect, 49,
St. Petersburg 197101, Russian Federation <br> </span> </div> <div> </div>
<br> <br> <span> DOI: &nbsp;</span> <span> 10.1016/j.dss.2013.07.008 </span>
<br> <span> Document Type: &nbsp;</span> <span> Article </span> <br> <span>
Source: &nbsp;</span> <span> Scopus </span> <br> </div> <span> 
<br> </span> <br> </div>

```

Фрагмент экспортируемого файла из базы данных Web of Science:

```

<tr> <td valign="top"> AU </td>< td> Mouromtsev, D <br> Pshenichny, C <br>
Yakovlev, A </td> </tr> <tr> <td valign="top"> AF </td> <td> Mouromtsev, Dmitry
<br> Pshenichny, Cyril <br> Yakovlev, Anthony </td> </tr><tr> <td valign="top">
TI </td> <td> Semantic and structural delineation of market scenarios by
the eventbush method </td> </tr> <tr> <td valign="top"> DE </td>
<td> Event bush method; Scenario; Consulting; Market; Knowledge engineering;
Expert </td> </tr> <tr> <td valign="top"> AB </td> <td> Considered is the
retrospective application of a new method of knowledge engineering,
the event bush, to a real collision that took place in the North-American market
of cool sparkling drinks in the 1980s... </td> </tr> <tr> <td valign="top">
C1 </td> <td> [Mouromtsev, Dmitry; Pshenichny, Cyril; Yakovlev, Anthony]
Natl Res UnivInformat Technol Mech & Opt, Intellectual Syst Lab,
St Petersburg197101, Russia. </td> </tr> <tr> <td valign="top"> SN </td>
<td> 0167-9236 </td> </tr> <tr> <td valign="top"> EI </td> <td> 1873-5797 </td>
</tr> <tr> <td valign="top"> PD </td> <td> JAN </td> </tr> <tr> <td valign="top">
PY </td> <td> 2014 </td> </tr> <tr> <td valign="top"> VL </td> <td> 57 </td> </tr>
<tr> <td valign="top"> BP </td> <td> 34 </td> </tr> <tr>
<td valign="top"> EP </td> <td> 41 </td> </tr> <tr> <td valign="top">
DI </td> <td> 10.1016/j.dss.2013.07.008 </td> </tr> <tr> <td valign="top">
PG </td> <td> 8 </td> </tr> <tr> <td valign="top"> WC </td>
<td> Computer Science, Artificial Intelligence; Computer Science,
Information Systems; Operations Research & Management Science </td>
</tr> <tr> <td valign="top"> UT </td> <td> WOS:000330909700004 </td> </tr>

```

Приведенные фрагменты экспортируемых файлов демонстрируют необходимость предварительного анализа и обработки представленной информации. Название публикации объединено со ссылкой на данную публикацию, сведения о выходных данных публикации представлены в виде одной строки, и необходим инструмент для извлечения сведений о публикации. Сведения о публикации, полученные из базы данных Web of Science, являются структурированными и значительно упрощают задачу обработки данных. В качестве модуля предварительной обработки был разработан достаточно гибкий инструмент, который позволяет производить предварительные настройки обработки конкретных сведений из заданной информационной системы. Модуль также включает в себя инструмент разбора строки согласно заданным шаблонам. Следующим

шагом является непосредственная обработка предварительно структурированных данных. Данная обработка содержит два глобальных модуля — модуль обработки сведений о публикациях и модуль обработки сведений об авторских коллективах. Модуль обработки сведений о публикации анализирует сведения, уже существующие в центральной базе данных университета, в случае выявления дубликатов происходит анализ полноты существующих данных и их обогащения при необходимости. Модуль идентификации авторов анализирует сведения об авторах, выявляет наиболее вероятных авторов на основании уже существующих сведений в БД университета об авторских коллективах. В данной работе к научным коллективам авторы относят:

- Соавторы. Система анализирует авторские связи, что позволяет выявить регулярность печати в соавторстве.
- Сотрудники одного подразделения (кафедр).
- Научное руководство студентами и аспирантам.
- Научные лаборатории.
- Участники проектов.

Так как обработка сведений о публикации включает не только обработку русскоязычных, но и обработку иностранных публикаций, необходимо производить перевод фамилий для дальнейшего анализа авторских коллективов. В качестве основного правила транслитерации было использовано «OVIR of Russia regulations». В информационной системе университета предусмотрена возможность хранения различных вариантов транслитерации фамилии авторов, что позволяет использовать любые правила транслитерации и их комбинации.

В связи с тем, что существуют различные методы транслитерации, не всегда возможно однозначно получить русскоязычное написание фамилии авторов. С учетом данного фактора возможно также и неоднозначное определение потенциальных авторов из базы физических лиц университета. В случае такой неоднозначности модуль идентификации авторов на основе анализа авторских коллективов выявляет данного соавтора (см. рис. 2а).

Рассмотрим пример, представленный на рис. 2 а более детально. Обозначим сотрудников университета как Ч x1.x2, где x1 — порядковый номер возможного написания, а x2 — порядковый номер сотрудника для данного описания. У публикации на английском языке указаны два автора: Dzerzhayskaya T.A., Varenikov D.A. Для того чтобы идентифицировать сотрудников, являющихся авторами данной публикации необходимо по иностранному написанию фамилии, имени и отчества найти в базе данных соответствующих сотрудников. Для данного примера были найдены следующие совпадения:

1. Автор 1 — Dzerzhayskaya T.A. Для данного автора были найдены следующие схожие написания:
 - Написание 1 — Dzerzhayskaya T.A. Данное написание указано у двух сотрудников:
 - Ч 1.1 — Державская Т.А.
 - Ч 1.2 — Державская Т.А.
 - Написание 2 — Dziarzhayskaya T.A. Данное написание определено на основании анализа иностранного написания фамилий авторов хранящихся в системе у одного сотрудника:
 - Ч 2.1 — Дзержавская Т.А.

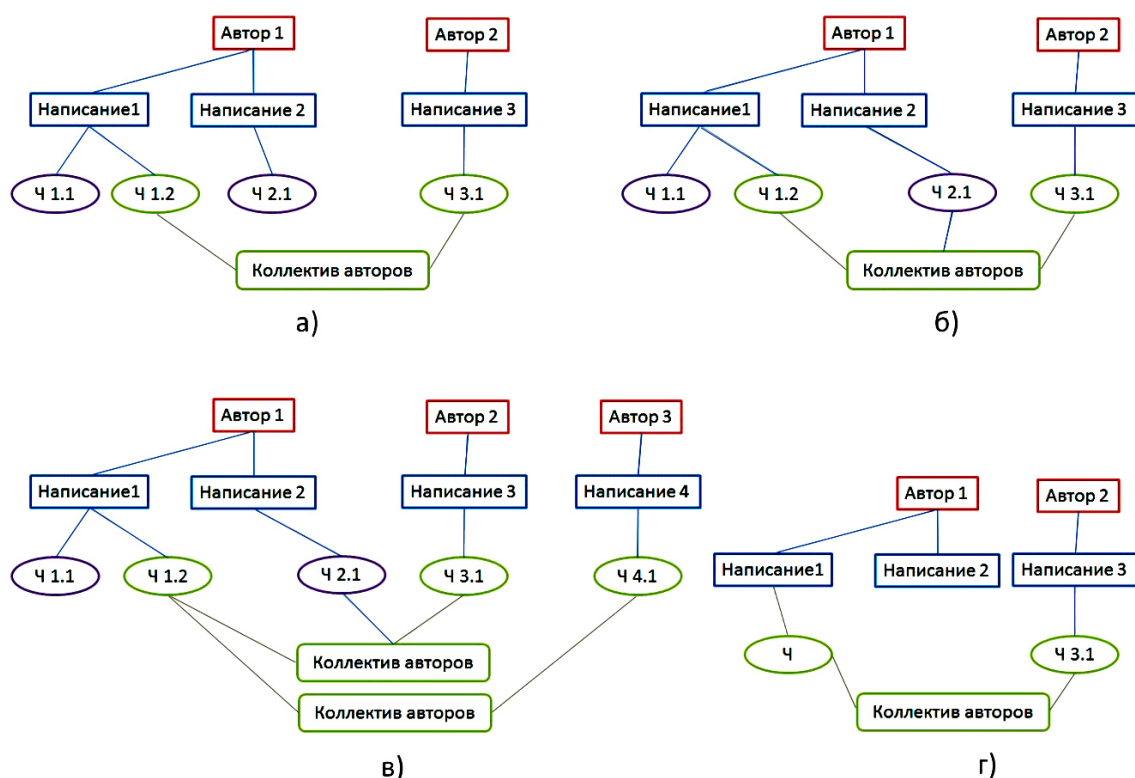


Рис. 2. Подход к идентификации авторов

2. Автор 2 — Varenikov D.A. Для данного автора было найдено одно написание:

- Написание 3 — Varenikov D.A.
- Ч 3.1 — Вареников Д.А.

Таким образом, однозначно определить связь Автора 1 с сотрудником невозможно, в отличие от Автора 2 для которого была найдена только одна связь с сотрудником Ч 3.1. Для того чтобы определить Автора 1 используется анализ авторских коллективов. С помощью проведенного анализа удалось определить, что из сотрудников Ч 1.1, Ч 1.2, Ч 2.1 только сотрудник Ч 1.2

Кроме того, возможен вариант неоднозначного определения соавтора после анализа авторских коллективов (см. рис. 2б) и дополнительных сведений об авторах, в этом случае система оставляет данного автора нераспознанным и формирует подсказку для специалиста, который в дальнейшем будет обрабатывать публикацию. Чем больше авторов приведено в публикации и чем полнее они описаны, тем точнее происходит идентификация авторов на основе авторских коллективов (см. рис. 2в). На рис. 2г показан пример неоднозначного определения автора после транслитерации. В данном примере идентификация соавтора происходит только после анализа авторского коллектива и обработки специалистами публикации на основании рекомендаций, представленных системой. Данный пример демонстрирует наполнение авторского профиля различными вариантами транслитерации его фамилии, что в дальнейшем позволяет идентифицировать его более точно. После анализа авторского коллектива данные о публикации

попадают в центральную базу научно-практических результатов. В таблице 1 приведен пример представления сведений о публикации в информационной системе университета.

Таблица 1. Представление описания сведений публикации в информационной системе университета

Поля	Значение
Год	2014
Тип результата	Статья
Наименование результата	Semantic and structural delineation of market scenarios by the event bush method
Язык	английский
Наименование издания	Decision Support Systems
Номер издания	1
Том	57
Номера страниц	34–41
Объем в страницах	
Вид материала	печатный текст
Издательство	Elsevier
Место издания	ELSEVIER SCIENCE BV, PO BOX 211, 1000 AE AMSTERDAM, NETHERLANDS
ISSN	0167-9236
e-ISSN	1873-5797
ISBN	
URL	http://www.sciencedirect.com/science/article/pii/S0167923613001930
URL Scopus	http://www.scopus.com/record/display.url?eid=2-s2.0-84892371567&origin=resultslist&sort=plf-f&src=s&st1=10.1016%2Fj.dss.2013.07.008&sid=38DF36687A8F9BB38EBC66A50378B090.WIW7NKKC52nnQNxjqAqrla%3a3230&sot=b&sdt=b&sl=30&s=DOI%2810.1016%2Fj.dss.2013.07.008%29&relpos=0&relpos=0&citeCnt=0&searchTerm=DOI%2810.1016%2Fj.dss.2013.07.008%29
URL WOS	http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=15&SID=W1fBp6TYYXoYeJAEXGf&page=1&doc=1
URL РИНЦ	http://elibrary.ru/item.asp?id=21862937
DOI	10.1016/j.dss.2013.07.008
WOS ID	000330909700004
Аффилиация	Да
Библиографическая запись	Mouromtsev, D.I. Semantic and structural delineation of market scenarios by the event bush method / D.I. Mouromtsev, C.A. Pshenichny, A.V. Yakovlev // Decision Support Systems. - 2014. - Vol. 57. - № 1. - Pp. 34-41
Библиографическая ссылка	Mouromtsev D.I., Pshenichny C.A., Yakovlev A.V. Semantic and structural delineation of market scenarios by the event bush method. Decision Support Systems. 2014. Vol. 57. No. 1. pp. 34-41.
Выходные данные	Mouromtsev D.I., Pshenichny C.A., Yakovlev A.V. Semantic and structural delineation of market scenarios by the event bush method // Decision Support Systems - 2014, Vol. 57, No. 1, pp. 34-41

В таблице 2 представлена информация об авторах публикации — иностранное описание, аффилиация, а также последовательность авторов в публикации.

Таблица 2. Представление информации об авторах

Авторы	Иностранное описание	Аффилиация	Посл.
Муромцев Дмитрий Ильич	Mouromtsev Dmitry I	да	1
Пшеничный Кирилл Анатольевич	Pshenichny Cyril A	да	2
Яковлев Антон Викторович	Yakovlev Anthony V	да	3

Полученные сведения могут использоваться для оперативных отчетов авторами публикаций, кафедрами, научными лабораториями, университетом. С целью повышения качества данных о публикациях, в университете организован отдел библиотеки, который осуществляет проверку всех существующих в центральной базе сведений о публикациях.

Работоспособность предложенного подхода была проверена на загрузке сведений о публикациях сотрудников Университета ИТМО за 2014 год из баз данных Web of Science и Scopus. Анализ полученных данных совместно с сотрудниками отдела библиотеки показал возможность использования разработанных модулей в информационной системе университета.

4. ЗАКЛЮЧЕНИЕ

В работе предложен подход к автоматизации поступлений сведений о научно-практических результатах в информационную систему университета. Его реализация позволила оптимизировать трудозатраты сотрудников, преподавателей и исследователей по внесению сведений о публикациях, значительно был снижен процент дублирующейся информации и ошибок персонального ввода авторами сведений о публикациях. Полнота полученных данных позволила также оптимизировать учет публикаций специалистами и, как следствие, повысить качество отчетных данных. Продолжение данной работы будет направлено на загрузку сведений о других результатах научно-исследовательской деятельности, а также разработку сервисов поиска и загрузки сведений о публикациях из внешних баз данных пользователями информационной системы в реальном времени.

Список литературы

1. Вареников Д.А., Попова И.А., Шлей М.Д. Вопросы централизованного хранения результатов проектной деятельности // Информационная среда вуза XXI века: материалы V Международной научно-практической конференции, 2011. С. 153–155.
2. Вареников Д.А. Информационные и технологические решения системы централизованного хранения результатов научно-практической деятельности // Сборник работ аспирантов НИУ ИТМО, победителей конкурса грантов правительства Санкт-Петербурга, 2013. С. 26–35.
3. Казин Ф.А., Биккулов А.С., Зленко А.Н., Тойвонен Н.Р., Попова И.А., Шлей М.Д., Вареников Д.А. Система поддержки проектной деятельности в Университете ИТМО // Инновации, 2014. № 8(190). С. 77–83
4. Мазов Н.А., Гуреев В.Н. Проблемы инентификации метаданных в наукометрических базах данных Web of Knowledge, Scopus и РИНЦ на примере профилей авторов // 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. М.: Изд-во ГПНТБ России, 2012. С. 1–4 / <http://www.gpntb.ru/win/interevents/crimea2012/disk/124.pdf> (дата обращения: 20.04.2015).

5. Наукометрические базы данных [электронный ресурс]. Режим доступа: <http://pspu.ru/university/biblioteka/prepodavatelam/indeksy-nauchnogo-citirovaniya>, свободный. Загл. с экрана (дата обращения: 20.04.2015).
6. *Ефимов М.Н., Шлей М.Д., Вареников Д.А.* Метод определения рекомендаций для пользователей информационной системы на основе их научных интересов и активности // Сборник материалов конференции «Научно-образовательная информационная среда XXI века» 2014. С. 71–73.
7. *Попова И.А., Вареников Д.А., Тойвонен Н.Р.* Информационные системы для научных исследований // Сборник научных статей. Труды XV Всероссийской объединенной конференции «Интернет и современное общество», 2012. С. 156—159.
8. *Мазов Н.А., Гуреев В.Н.* Идентификация библиографических метаданных научных публикаций в различных базах данных: проблемы и решения // Материалы 8-й Междунар. конф. посвящ. 60-летию ВИНТИ РАН «Актуальные проблемы информационного обеспечения науки, аналитической и инновационной деятельности», «НТИ – 2012». 28–30 ноября 2012 г., ВИНТИ РАН, г. Москва. М., 2012. С. 123–124 / <http://www.viniti.ru/download/russian/prog8.pdf> (дата обращения: 20.04.2015).
9. *Коляда А.С., Гогунский В.Д.* Автоматизация извлечения информации из наукометрических баз данных // Управління розвитком складних систем, 2013. № 16. С. 96–99.
10. *Пинжин А.Е.* Применение вероятностного алгоритма соединения записей для исключения дублирования информации в корпоративной базе данных // Известия Томского политехнического университета. Томск, 2006. № 7. С. 111–116.

APPROACHES TO AUTOMATION OF DATA PROCESSING IN SCIENTOMETRIC DATABASES

Varenikov D.A., Mouromtsev D.I., Shley M.D.

Abstract

In this paper authors present the results of the development of an approach to automate the loading of data from scientometric databases into the University knowledge base. The basic problem of processing information about publications is described. Particular attention is paid to the identification when the authors of this publication are the university employees. The implementation of the proposed approach has allowed the university to optimize labor costs of employees to fill the University knowledge base, as well as improve the quality of the stored data.

Keywords: *scientometric database, automation, publishing, authors, co-authors, Web of Science, Scopus, Russian Science Citation Index, identification of the authors, affiliation, information system.*

Вареников Денис Анатольевич,
заместитель начальника отдела
департамента информационных
технологий отдела информационной
поддержки проектной деятельности
Университета ИТМО,
varenikovda@gmail.com

Муромцев Дмитрий Ильич,
кандидат технических наук, доцент,
руководитель международной
лаборатории «Интеллектуальные методы
обработки информации и семантические
технологии» Университета ИТМО,
mouromtsev@mail.ifmo.ru

Шлей Михаил Дмитриевич,
кандидат технических наук, начальник
отдела департамента информационных
технологий отдела информационной
поддержки проектной деятельности
Университета ИТМО,
mikhail.shlei@gmail.com

©

Наши авторы, 2015.

Our authors, 2015.