УДК 025.43
DOI 10.33186/1027-3689-2019-3-75-84

Н. С. Солошенко

Всероссийский институт научной и технической информации РАН

Глубокое тематическое индексирование документов в системе комплектования информационного центра

Представлены основные технологические этапы глубокого тематического индексирования и методы его использования в комплектовании входного документного потока информационно-аналитического центра. Рассмотрены проблемы тематической атрибуции журналов в зарубежных политематических информационно-аналитических ресурсах при индексировании изданий в целом. Показаны преимущества формирования неограниченного числа предметных рубрик при индексировании документов и создании множественных тематических профилей сериальных изданий, свидетельствующих о возрастающем количестве междисциплинарных исследований.

Проведено условное сравнение тематического распределения статей 2016–2017 гг. по химическим тематикам в массиве российских журналов, обязательных для отражения в БД ВИНИТИ РАН, и в журнальном массиве БД Scopus. Анализ документных массивов БД Scopus проведён с помощью инструментария аналитического ресурса Scimago. Выявлено существенное несоответствие распределения документных массивов по аналогичным тематикам в обеих БД, что обусловлено разными принципами индексирования. Показано, что документное индексирование позволяет службе комплектования информационно-аналитического центра следить за информационным наполнением различных тематических направлений и обеспечивать полноту результирующих выборок документов при подготовке информационных продуктов.

Ключевые слова: тематическая атрибуция, тематическое индексирование изданий, тематическое документное индексирование, информационный центр, реферативно-аналитические базы данных, тематический профиль журнала.

UDC 025.43
DOI 10.33186/1027-3689-2019-3-75-84

Nataliya Soloshenko
All-Russian Institute for Scientific and Technical Information of the Russian Academy of Sciences, Moscow, Russia

Advantages of document deep thematic indexing for the information center acquisitions service

The paper presents the main technological stages of deep thematicindexing and its benefitsfor the information center acquisitions service. The most common problems of journal thematic attribution as a whole in global citation indexes are considered. It is argued that deep document indexing leads to an unlimited number of document subject headings thus creating serials multiple thematic profiles that represent an increasing number of interdisciplinary studies. The specific examples demonstrate the dynamic links formation between the publication and a certain set of thematic headings, which makes it possible to identify productive publications with different basic profiles for specific subject areas.

A comparative study of 2016–2017 chemistrydocument arrays for VINITI RAS and Scopus journals selections was conducted. The Scopus journals analysis was carried out using the Scimagotools. In the Scopus database, the most representative thematic categoryisdetermined as "Chemistry, miscellaneous", which includes both polythematic and highly specialized publications.The maximum flow of Russian editionsarticles is noted in the field of physical chemistry with a significant number of productive journals with a basic physical profile.

This model of deep thematic indexing allows the information center acquisitionsservice to monitor the content of various subject areas and ensure the completeness of the resulting documentsetsfor theinformation products.

Keywords: thematic attribution, publication thematic indexing, document thematic indexing, information center, citation databases, publication thematic profile.

A typical problem with polythematic database is the lack of mutual correspondence between thematic categories/rubrics. For example, the Herald of the Russian Academy of Sciences in the WoS CC is referred to multidisciplinary research. In Scopus, this journal is indexed as “Cultural Studies” and “Political Science and International Relations”. The original and translated versions of Russian journals in foreign databases correspond to different subject areas or categories. "News of the Russian Academy of Sciences. Biological series” and “Molecular genetics, microbiology and virology” in the Scopus database the original versions are attributed to medicine, while their translated versions are related to biochemistry, molecular biology, microbiology, genetics. While formation our DB VINITI, we carry out documentary thematic indexing at different stages of processing. At the first stage, information is entered in accordance with the formal inprints or

on the publication's website. Then, at the stage of the primary thematic profiling, specialists assign thematic attributes to specific documents in fairly broad areas of research relevant to the subject editions. Each document has a unique identification code, and information about it goes to the Unified Technological Database at all stages of its processing. Thus, in this database, document can be allocated both at the analytical level of a separate publication, and at the level of the publication as a whole. Thus, thematic documentary indexing allows the service of picking the input stream of the information-analytical center: create multiple thematic profiles of sources; to identify publications with a significant number of articles marked up / reflected in the database, indexed by a specific heading, for compiling lists of journals that are “mandatory” for reflection in the corresponding issue of the database of the VINITI RAS; to identify publications that, for various reasons, have ceased to be specialized for a specific subject matter; follow the content of various thematic areas.

Тематическая атрибуция документов – общая задача для информационно-аналитических ресурсов. Практически все зарубежные политематические реферативно-аналитические базы данных и в частности наиболее распространённые из них – *Web of Science Core Collection (WoS CC)* и *Scopus* – индексируют периодические источники в целом. При этом зачастую отсутствует взаимное соответствие тематических категорий/рубрик *Journal Citation Reports* (*JCR*), *WoS CC* и *Scopus* [**1**, **2**]. Так, например, «Вестник Российской академии наук» *(Herald of the Russian academy of Sciences)* в БД (*WoS CC*) отнесён к мультидисциплинарным исследованиям. В БД *Scopus* этот журнал индексируется по категориям «Культурология» и «Политология и международные отношения» (табл. 1). Пример с британским журналом *«Advanced Functional Materials»* так же наглядно демонстрирует разные подходы к тематическому индексированию изданий у предметных экспертов этих ресурсов.

Таблица 1

Тематическая атрибуция англоязычных журналов в БД WoS CC и Scopus

Наименование журнала	**Тематические категории WoS CC**	**Тематические категории Scopus**
Herald of the Russian academy of Sciences	Multidisciplinary Sciences	Cultural Studies; Political Science and International Relations
Advanced Functional Materials (GB)	Chemistry, Multidisciplinary – Scie; Chemistry, Physical – Scie; Nanoscience & Nanotechnology – Scie; Materials Science, Multidisciplinary – Scie; Physics, Applied – Scie; Physics, Condensed Matter – Scie	Electrochemistry; Biomaterials; Electronic, Optical and Magnetic Materials; Condensed Matter Physics

Кроме того, в ряде случаев оригинальным и переводным версиям российских журналов в зарубежных БД соответствуют разные предметные области или категории [**1**]. Так, оригинальные версии журналов «Известия РАН. Серия биологическая» и «Молекулярная генетика, микробиология и вирусология» в БД *Scopus* приписаны к медицине [**3**], в то время как их переводные версии – к биохимии, молекулярной биологии, микробиологии, генетике (табл. 2).

Таблица 2

Тематическая атрибуция оригинальных и переводных версий журналов в БД Scopus

Наименование журнала	Тематические категории Scopus (оригинальная версия)	Тематические категории Scopus (переводная версия)
«Молекулярная генетика, микробиология и вирусология»	Medicine (miscellaneous)	Genetics; Molecular Biology; Microbiology; Virology; Infectious Diseases
«Известия РАН. Серия биологическая»	Medicine (miscellaneous)	Agricultural and Biological Sciences (miscellaneous); Biochemistry, Genetics and Molecular Biology (miscellaneous)

Отсутствие единообразия в индексации изданий в зарубежных БД требует дополнительного применения релевантных ключевых слов при поиске по конкретной тематической категории.

В отличие от зарубежных политематических информационных ресурсов в БД ВИНИТИ РАН осуществляется документное тематическое индексирование, которое проводится на разных стадиях обработки входного потока для подготовки информационных продуктов.

На первом этапе обработки журналов в Базовый массив сериальных изданий Автоматизированной системы комплектования регистрации ВИНИТИ РАН вносится информация о профильных тематических областях издания (химия, физика, биология и др.) в соответствии с формальными признаками, указанными в печатном экземпляре или на сайте издания.

Далее, на этапе первичного тематического профилирования документов специалисты, работая с изданием на аналитическом уровне, присваивают тематические признаки конкретным документам по достаточно широким направлениям исследований, соответствующим профилям работы отделов научной информации (предметных редакций): автоматика и радиоэлектроника, биологические и медико-биологические дисциплины, химия и химические технологии, физика, астрономия и др.

При поступлении документов в отделы научной информации проводится их тематическое индексирование редакторами конкретных выпусков Реферативного журнала/БД ВИНИТИ РАН в соответствии с Рубрикатором отраслей знаний ВИНИТИ РАН, построенным на основе углубления Государственного рубрикатора научно-технической информации (ГРНТИ) (3-го уровня) по мере потребности отдельных отраслей до 9-го уровня [**4**]. Отдельный документ может быть проиндексирован разными редакциями и иметь неограниченное число предметных рубрик.

Каждый документ имеет уникальный идентификационный код, сведения о нём поступают в Единую технологическую БД (ЕТБД) на всех стадиях обработки. При этом отдельный документ привязан в конкретному выпуску соответствующего издания. Таким образом, в ЕТБД собираются документные массивы, которые могут быть выделены как на аналитическом уровне отдельной публикации, так и на уровне издания в целом.

Инструментарий ЕТБД позволяет отслеживать документный поток по различным тематическим направлениям в режиме реального времени.

Для представления возможностей глубокого тематического индексирования проанализирован массив российских сериальных изданий, которые имеют в своих тематических профилях рубрики ГРНТИ с верхним индексом «31. Химия» (всего 305 наименований) и обязательно должны быть отражены в БД по химии. Основными рубриками второго уровня, по которым индексируются статьи, являются: 31.15 Физическая химия; 31.17 Неорганическая химия. Комплексные соединения; 31.19 Аналитическая химия; 31.21 Органическая химия; 31.23 Биоорганическая химия. Природные органические соединения и их синтетические аналоги; 31.25 Химия высокомолекулярных соединений.

Из первоначального массива были отобраны издания, имеющие рубрики с верхним индексом «31. Химия» в качестве первых трёх по количеству отражённых в 2016–2017 гг. документов (всего 118 наименований), например: «Коллоидный журнал» проиндексирован следующими рубриками: 31.15, 61.57, 61.13, 61,71 и включен в анализируемый массив, так как содержит в первой тройке рубрику 31.15.

В соответствии с законом Брэдфорда в полученной выборке выделены три группы изданий в порядке убывания количества отображённых в них профильных документов, проиндексированных в БД ВИНИТИ химическими рубриками (каждая группа изданий содержит примерно $^1/_3$ полученного документного массива): издания с высокой продуктивностью (ядерные); издания со средней продуктивностью; малопродуктивные издания. В то же время закон Брэдфорда невозможно применить в полной мере к этой совокупности журналов, представляющих не эмпирическую, а экспертную выборку изданий.

В категорию высокопродуктивных химических журналов вошли всего семь изданий: «Известия Российской академии наук (РАН). Сер. Химическая», «Журнал общей химии», «Журнал физической химии», «Журнал органической химии», «Журнал неорганической химии», «Журнал структурной химии», «Вестник Казанского государственного технологического университета».

В группу со средней продуктивностью включены 18 журналов. Выявлены 94 малопродуктивных издания. В эту категорию попали такие авторитетные журналы, как «Успехи химии», «Вестник МГУ. Сер. 2. Химия», «Высокомолекулярные соединения. Сер. А, В, С». Это свидетельствует о том, что продуктивность журналов никак не связана с научной ценностью их публикаций и является только количественным показателем.

Аналогичное деление на категории высоко-, средне- и малопродуктивных изданий было проведено в полученной выборке журналов по каждой основной рубрике второго уровня. Тематическое индексирование документов устанавливает динамические связи между изданием и его тематическим профилем; в продуктивные категории отдельной рубрики могут попасть журналы, основные профили которых не совпадают с этой тематикой.

Так, например, в группу продуктивных журналов по физической химии входят такие издания, как «Физика твёрдого тела» и «Журнал технической физики», профили которых в различных информационных ресурсах представлены в табл. 3.

Таблица 3

Тематические профили продуктивных журналов по физической химии в различных БД

Наименование журнала	Тематические рубрики ВИНИТИ (1–3)	РИНЦ	Тематические категории Scopus (переводная версия)
«Физика твёрдого тела»	29.19 Физика твёрдых тел; 31.15 Физическая химия; 29.31 Оптика	29.19 Физика твёрдых тел	Materials Science: Electronic, Optical and Magnetic Materials; Physics and Astronomy: Condensed Matter Physics
«Физика и техника полупроводников»	29.19 Физика твёрдых тел 31.15 Физическая химия 29.31 Оптика	29.19.31 Полупроводники	Physics and Astronomy: Condensed Matter Physics; Materials Science: Electronic, Optical and Magnetic Materials; Physics and Astronomy: Atomic and Molecular Physics, and Optics

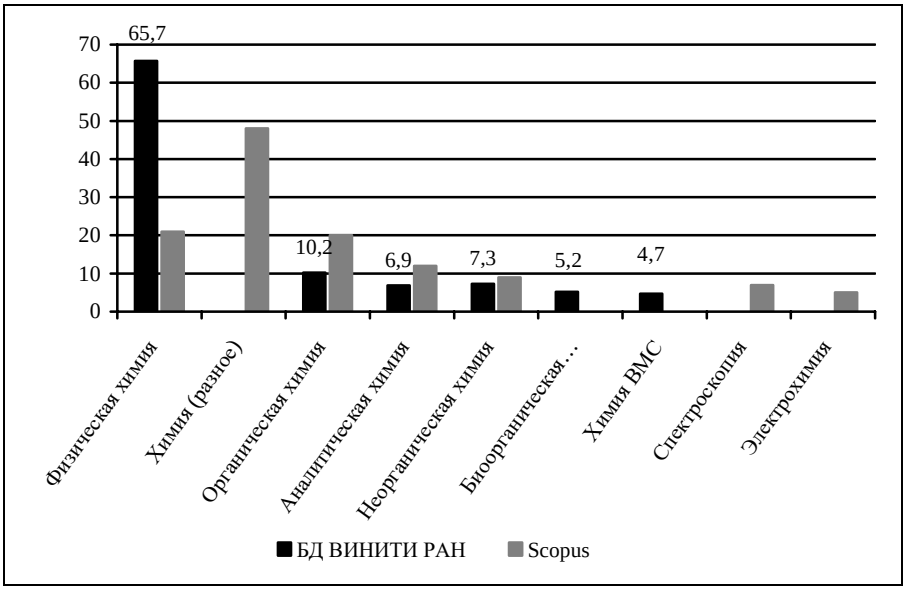
Окончание таблицы 3

Наименование журнала	Тематические рубрики ВИНИТИ (1–3)	РИНЦ	Тематические категории Scopus (переводная версия)
«Физика металлов и металловедение»	53.49 Металловедение; 29.19 Физика твёрдых тел; 31.15 Физическая химия	29.19 Физика твёрдых тел; 53.49 Металловедение	Materials Science: Materials Chemistry; Physics and Astronomy: Condensed Matter Physics
«Журнал технической физики»	29.19 Физика твёрдых тел; 29.27 Физика плазмы; 31.15 Физическая химия	29.00 Физика	Physics and Astronomy (miscellaneous)

Журналы «Физика твёрдого тела» и «Физика и техника полупроводников» и в БД ВИНИТИ РАН, и в РИНЦ проиндексированы рубрикой 29.19 «Физика твёрдых тел», аналогичная рубрика *«Condensed Matter Physics»* есть у переводной версии журнала *«Solid State Physics»* и *«Semiconductors»* в БД *Scopus*. Почти все тематические рубрики журнала «Физика металлов и металловедение» совпадают во всех трёх ресурсах в отличие от «Журнала технической физики», включая его переводную версию *«Technical Physics»*, притом, что тематика «Физическая химия» не указана для всех приведённых выше изданий ни в РИНЦ, ни в БД *Scopus*.

Из-за различных принципов тематического индексирования возможно лишь условное сравнение распределения документных массивов даже по аналогичным тематикам в БД ВИНИТИ РАН и БД *Scopus*. Для выполнения этой задачи были выделены массивы статей 2016–2017 гг., проиндексированных основными рубриками по химии второго уровня ГРНТИ из анализируемой выборки журналов, обязательных для отражения в БД ВИНИТИ РАН.

Исследован и журнальный массив по химии в информационном ресурсе *SciMago* (Университет Гранады, Испания) [**3**], анализирующем данные из БД *Scopus*. В результате были выделены подборки изданий с их документными массивами за 2017 г. по тематикам, аналогичным рубрикам ГРНТИ. Проведено сравнение сходных тематических документных массивов в обеих БД в отношении к общим массивам статей по химии (см. рис.).



Относительные доли отдельных тематик
в общем объёме журнальных статей по химии (%)

Данные, приведённые на рисунке, свидетельствуют о том, что максимальное количество статей, проиндексированных рубриками ГРНТИ с верхним уровнем «31. Химия» из обязательных для БД ВИНИТИ РАН журналов, включено в тематику «Физическая химия» (66%). Это можно объяснить отмеченным выше отсутствием жёсткой связи между изданием и определённым набором рубрик, что позволяет учитывать документы из источников с множественными тематическими профилями.

В БД *Scopus* максимальный журнальный массив входит в категорию «Химия (разное)», которая включает издания, охватывающие почти все области химии, например переводную версию журнала «Известия Академии наук. Сер. Химическая», и узкоспециализированные издания, например переводную версию журнала «Координационная химия». Кроме того, в БД *Scopus* тематика «Электрохимия» выведена отдельно, в то время как в ГРНТИ она является рубрикой третьего уровня в составе «Физической химии».

Значительная разница в относительных объёмах журнальных статей по органической химии в БД ВИНИТИ РАН и БД *Scopus* объясняется в большей степени тем, что в БД *Scopus* в эту тематику включаются издания, отнесённые в ГРНТИ к рубрикам «31.23. Биоорганическая химия» и «31.25. Химия высокомолекулярных соединений». Следует ещё раз отметить: разные принципы индексирования не допускают полноценного сравнения тематических документных массивов БД ВИНИТИ РАН и зарубежных ресурсов.

Одна из особенностей системы глубокого тематического индексирования – возможность формировать множественные тематические профили изданий (что свидетельствует о возрастающем количестве междисциплинарных исследований) и обеспечивать полноту результирующих выборок документов при подготовке тематических обзоров и других аналитических информационных продуктов.

Таким образом, тематическое документное индексирование позволяет службе комплектования входного потока информационно-аналитического центра:

формировать множественные тематические профили источников;

выявлять издания со значительным количеством размеченных/отражённых в БД статей, проиндексированных конкретной рубрикой, для составления перечней журналов, обязательных для отражения в соответствующем выпуске БД ВИНИТИ РАН;

выявлять издания, в силу различных причин переставшие быть профильными для конкретной тематики;

следить за информационным наполнением различных тематических направлений.

СПИСОК ИСТОЧНИКОВ

1. **Зибарева И. В., Солошенко Н. С.** Российские журналы в глобальных информационно-аналитических ресурсах // Вестн. РАН. – 2016. – Т. 86, № 9. – С. 824–838.

Zibareva I. V., Soloshenko N. S. *Rossiyskie zhurnaly v globalnyh informatsionno-analiticheskih resursah // Vestn. RAN. – 2016. – T. 86, № 9. – S. 824–838.*

2. **Глушановский А. В., Каленов Н. Е.** Некоторые сравнительные характеристики баз данных Scopus и Web of Science // Информация и инновации. – 2016. – № 2. – С. 15–19.

Glushanovskiy A. V., Kalenov N. E. *Nekotorye sravnitelnye harakteristiki baz dannyh Scopus i Web of Science // Informatsiya i innovatsii. – 2016. – № 2. – S. 15–19.*

3. **Scimago** Journal & Country Rank (SCImago). – URL: http://www.scimagojr.com.

4. **Рубрикатор** отраслей знаний ВИНИТИ РАН (РВИНИТИ РАН). – Режим доступа: http://www.viniti.ru/products/classification-systems/rubricator-viniti.
Rubrikator *otrasley znaniy VINITI RAN (RVINITI RAN).*

Nataliya Soloshenko, *Cand. Sc. (Pedagogy), Acquisitions Department, All-Russian Institute for Scientific and Technical Information (VINITI), Russian Academy of Sciences;*
solns@viniti.ru
20, Usiyevicha st., 125190 Moscow, Russia