Терминологические словари в системе сопоставления библиографических классификаций[1]

В. Н. Белоозеров,
Всероссийский институт научной и технической информации РАН,
Москва, Россия

В докладе рассмотрены работы по созданию словарей терминов различных тематических разделов ГРНТИ, проводимые в рамках задания Минобрнауки РФ по сопоставлению классификаций в сфере научно-технической информации. Разработана лексикографическая система, состоящая из 63 отдельных словарей, по тематике главных подразделений ГРНТИ, покрывающая всё поле научно-технического знания. Лексика словарей основана на выделении терминов из наименований рубрик ГРНТИ и связанных с данной тематикой рубрик 10 других классификаций. Источниками определений и дополнительных терминов являются авторитетные энциклопедии и специальные словари, а также опыт индексирования научных публикаций ВИНИТИ, БЕН РАН и НПБУ. Определено место терминологических словарей при создании объединённой классификационно-словарной системы навигации по разнородным информационным ресурсам.

Разработка терминологических словарей по лексике классификационных систем научно-технической информации проводится в ВИНИТИ РАН в соответствии с соглашением с Министерством образования и науки Российской Федерации о реализации проекта «Сопоставление ГРНТИ с другими классификационными системами с целью совершенствования системы тематической кодификации НИР, НИОКР гражданского назначения. Формирование системы соответствий между различными классификаторами в сфере научно-технической информации»[2,3].

Потребность в словарях возникла у заказчика потому, что сопоставление классификаций рассматривалось как инструмент мониторинга, анализа содержания различных информационных ресурсов, который будут проводить администраторы, не являющиеся специалистами во всех научных областях. Для этого требуется не только формальное указание на соответствие рубрик, но и пояснение смысла того, чем наполнены эти рубрики. При этом словари должны быть специализированы по сопоставляемым рубрикам, чтобы целенаправленно раскрывать содержание определённой научной области. Тем самым, не ставилась задача создания одного всеобъемлющего словаря по всему универсуму знаний, а альтернативой могла быть разработка отдельных словарей по каждому элементу сопоставительной таблицы рубрик. Но поскольку в некоторых классификациях (УДК, ББК, МКИ) число рубрик в каждой превышает 100 000, стало очевидно, что такая работа практически невыполнима. Компромисс был найден в том, чтобы разрабатывать словари в соответствии с тематикой основных разделов ГРНТИ.

Всего в ГРНТИ насчитывается 63 раздела определённой научной тематики плюс ещё несколько обобщающих разделов, не выходящих за пределы понятий указанных тематических разделов. В 63 разделах ГРНТИ содержится почти 8000 рубрик, которые включают один или несколько терминов, подлежащих определению в словарях. Сопоставленные этим рубрикам ГРНТИ рубрики других одиннадцати классификаций многократно увеличили бы объём словарей и трудоёмкость работы. Приемлемое ограничение материала привело к тому, что было решено в словари включать, главным образом, термины, выделенные из рубрик ГРНТИ второго уровня и рубрик других классификаций, непосредственно соотносимых с главными разделами ГРНТИ. Разъяснение терминов этих рубрик верхнего уровня позволят администраторам составить представление

[1] Работа выполнена при финансовой поддержке Министерства образования и науки Российской Федерации, шифр проекта 2014-14-573-0024-001.

[2] Сопоставление ГРНТИ с другими классификационными системами с целью совершенствования системы тематической кодификации НИР, НИОКР гражданского назначения. Формирование системы соответствий между различными классификаторами в сфере научно-технической информации: Отчёт за 2014 г. по соглашению ВИНИТИ РАН и Минобрнауки России № 14.601.21.0001 / ВИНИТИ РАН; акад. Ю. М. Арский, И. Ю. Никольская, С. М. Гоннова и др. – М. 2014.

[3] Установление соответствий рубрик ГРНТИ рубрикам других систем классификации научной и технической информации / Антопольский А. Б., Белоозеров В. Н., Маркарова Т. С., Дмитриева Е. Ю. // Научно-техническая информация. Сер.1. Орг. и методика информ. работы. – 2015. № 3. – С. 3–19.

о содержании научной информации на достаточно обобщённом уровне, который вполне соответствует задачам управления научными работами и ресурсами.

Однако при любом применении рассматриваемой совокупности библиографических классификационных систем кроме сопоставления рубрик друг с другом присутствует необходимость сопоставлять рубрики с содержанием научных работ. А тематика научных работ отнюдь не всегда выражается терминами какой-либо классификации. Для того чтобы словари могли помочь и в процессе сопоставления с реальными научными работами, словари должны содержать в достаточной мере ту лексику, которая используется в текущих разработках. Такую лексику пришлось изъискивать в разных источниках и также включать в словари в режиме пополнения словников, составленных по терминологии классификационных рубрик.

В соответствии с идеей сосредоточиться на терминах основных рубрик 63 разделов ГРНТИ и сопоставленных им рубрик других классификаций на первом этапе работ была составлена сопоставительная таблица этих разделов рубрикам всех сопоставляемых классификаций. В этой таблице для каждого раздела ГРНТИ перечислены рубрики десяти классификаций, которые оптимальным образом соответствуют содержанию данного раздела ГРНТИ. Например, на рис. 1 показан фрагмент таблицы, содержащий перечень соответствий для раздела ГРНТИ ***27 Математика***.

В таблице указана степень соответствия рубрик, выписаны коды и наименования соответствующих рубрик, а также относящиеся к ним комментарии. Так для раздела ***27 Математика*** в большинстве других классификаций имеется точный эквивалент (обозначен символом ***Экв.***), но классификатор Web of Science делит математику на три раздела по не совсем понятным основаниям – фундаментальная математика, практическая математика и приложения математики в других науках. В иных случаях указаны рубрики охватывающие содержание раздела ГРНТИ и частично пересекающиеся в своей существенной части. Степень соответствия обозначена согласно таблице 1.

27 МАТЕМАТИКА

- ***УДК*** ***Экв.***: ***51*** Математика
- ***ББК*** ***Экв.***: ***22.1*** Математика
- ***ОЭСР*** ***Экв.***: ***1.1*** Mathematics
  Pure mathematics, Applied mathematics; Statistics and probability
- ***SCOPUS*** ***Экв.***: Mathematics
- ***WoS*** ***Ниже:*** Mathematics
  Mathematics covers resources having a broad, general approach to the field. The category also includes resources focusing on specific fields of basic research in Mathematics such as topology, algebra, functional analysis, combinatorial theory, differential geometry and number theory.
  Ниже: Mathematics, Applied
  Mathematics, Applied covers resources concerned with areas of mathematics that may be applied to other fields of science. It includes areas such as differential equations, numerical analysis, nonlinearity, control, software, systems analysis, computational mathematics and mathematical modeling. Resources that are concerned with mathematical methods and whose primary focus is on a specific non-mathematics discipline (except biology) such as psychology, history, economics etc., are covered in the MATHEMATICS, INTERDISCIPLINARY APPLICATIONS category. Resources focusing on mathematical biology are cjverd in MATHEMATICAL & COMPUTATIONAL BIOLOGY category.
  Ниже: Mathematics, Interdisciplinary Applications
  Mathematics, Interdisciplinary Applications includes resources concerned with mathematical methods whose primary focus is on a specific non-mathematics discipline (except biology) such as psychology, history, economics, etc. Resources that deal with mathematical biology are covered in the MATHEMATICAL AND COMPUTATIONAL BIOLOGY category. Resources that focus on specific mathematical topics such as differential equations, numerical analysis, nonlinearity, etc., are covered in the MATHEMATICS, APPLIED category
- ***ВАК*** ***Экв.:*** ***01.01.00*** Математика
- ***РНФ*** ***Экв.:*** ***01-100*** Математика
- ***РФФИ*** ***Экв.:*** ***01-100*** Математика
- ***ФАНО*** ***Асс.:*** Обработка и анализ больших массивов данных (Big Data)
- ***МПК*** ***Асс.:*** ***G06*** Обработка данных; вычисление; счет

Рис. 1 – Соответствия Разделу ГРНТИ ***27 Математика*** в других классификациях

Таблица 1 – Обозначения степени соответствия рубрик

Символ связи	Наименование связи	Описание смысловой связи рубрик
1	2	3
=	Эквивалентность (***Экв.***)	Тематика рубрик **совпадает**. Документ, отнесённый к одной рубрике, входит также в тематику другой.
<	Вышестоящая рубрика (***Выше***)	Сопоставленная рубрика имеет более **высокую** степень общности, чем рубрика ГРНТИ. Документ из рубрики ГРНТИ также входит в тематику сопоставленной рубрики, которая содержит также документы по иным темам.
>	Нижестоящая рубрика (***Ниже***)	Сопоставленная рубрика имеет более **низкую** степень общности, чем рубрика ГРНТИ. Документ из сопоставленной рубрики также входит в тематику данной рубрики ГРНТИ, которая, однако, содержит и документы по другим темам.
><	Ассоциация (***Асс.***)	Тематика рубрик **пересекается** в существенной части. Многие документы, отнесённые к каждой из рубрик, входят также в тематику другой рубрики.

Таблица соответствия верхнего уровня ГРНТИ всем классификациям даёт возможность обеспечить разработчиков словарей информацией об исходной лексике, подлежащей обработке. А также эта таблица задаёт направление и образец для разработки подробных таблиц соответствия всех нижних рубрик ГРНТИ.

Имея эту таблицу, каждый разработчик (всего разработчиков – восемь) выявлял из наименований и комментариев рубрик термины, отыскивал их определения в авторитетных источниках и заносил найденное в словарь в алфавитном порядке с указанием ссылки на источник. В качестве источников использовались главным образом энциклопедии и отраслевые энциклопедические словари, доступные в Интернете (как правило, на сайте «Академик.ру»). При этом обычно в словарь вносилась не полная энциклопедическая статья, а только её начальная часть, где давалось собственно определение понятия. Однако в некоторых случаях, особенно в областях общественных и гуманитарных наук приходилось давать всю или большую часть терминологической статьи, когда содержание понятия раскрывается не логическим определением, а многообразными связями с контекстом его бытования в той или иной науке. Заложенная в концепции нашей работы идея использовать массив стандартизованной терминологии оказалась малоэффективной. Тому имеется две причины. Во-первых, лексика, определяемая в стандартах на термины и определения, касается главным образом слишком частных понятий технических дисциплин, которые не соответствуют достаточно обобщённому уровню рассмотрения в классификациях. Во-вторых, определения в стандартах формулируются слишком формально, а для использования словарей в качестве пособия по раскрытию содержания научных работ требуется обычно не столько логическое определение понятия, сколько популярное толкование термина. Однако в ряде случаев в словари включались и статьи терминологических стандартов. В случаях, когда в разных источниках содержатся разные толкования термина, отличающиеся либо по содержанию понятия, либо по основаниям определения, то включались в одну терминологическую статью два или даже больше толкований, если это способствует более полному раскрытию значения термина.

На следующем этапе работы словари, составленные по лексике классификаций, пополнялись терминами текущих научных исследований. В качестве основного источника этой лексики были выбраны ключевые слова, которые индексируют содержание научных работ в базе данных ВИНИТИ. Для этого были привлечены индексаторы по всем тематическим разделам баз данных ВИНИТИ, которым было выдано задание указать для каждой рубрики ГРНТИ до пяти наиболее важных ключевых слов, приписанных к документам этой базы данных и сформулировать определения соответствующих понятий. Этот материал поступил на обработку ответственным исполнителям словарей, и соответствующие термины и определения были включены в словари после

редакционной обработки. Коды соответствующих рубрик ГРНТИ были включены в терминологические статьи разрабатываемых словарей. Общая структура словарных статей может быть проиллюстрирована следующим примером (см. рис. 2).

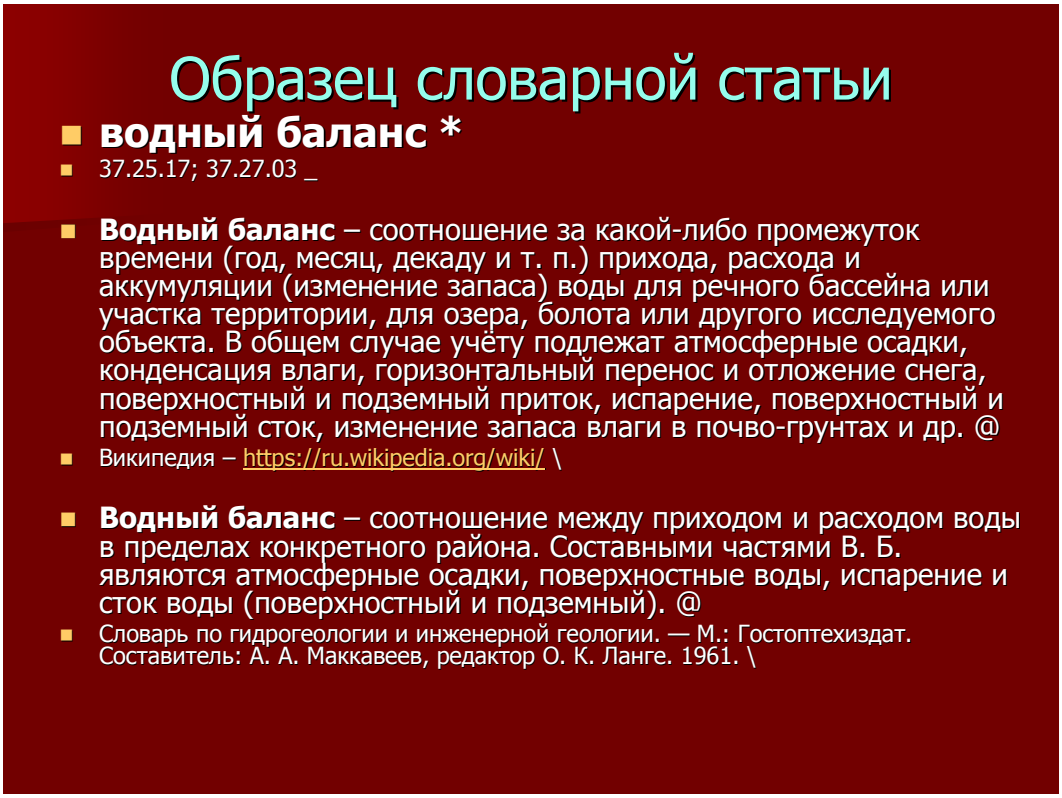

Рис. 2 – Словарная статья из словаря по водному хозяйству

С целью ввода словарей в компьютерные системы элементы статьи размечены служебными символами, которые выделяют в статье: заглавие (*), код ГРНТИ (_), толкование (@) и ссылки на источник (\ ).

К сожалению, в ВИНИТИ ведутся базы данных только по естественным и техническим наукам, и то не вполне исчерпывающе.

По гуманитарным и общественным наукам для пополнения словарей был привлечён тезаурус Научной педагогической библиотеки им. К. Д. Ушинского (НПБУ), на основании которого индексировалась база данных библиотеки, охватывающая в определённой мере тематику всех наук, по которым ведётся преподавание (исполнитель – Т. С. Маркарова).

Ключевые слова, используемые в Библиотеке по естественным наукам РАН (БЕН РАН), были внесены дополнительно к ключевым словам ВИНИТИ при пополнении словарей по физике, механике, астрономии, космическим исследованиям (исполнитель – Л. А. Верная), биологии, сельскому хозяйству, медицине (исполнитель – А. А. Ивановский).

Материалы Вычислительного центра РАН использованы в словарях по математике, кибернетике и вычислительной технике (исполнитель – Ю. О. Трусова). Материалы Института физики полупроводников СО РАН – в словарях по радиоэлектронике и связи (исполнитель – Н. Н. Шабурова).

Для многих разделов, где статистика ключевых слов была недостаточно представительной, в словари были включены термины всех трёх уровней рубрик ГРНТИ (исполнители – Ю. П. Косарская и А. Б. Антопольский).

Полученные от исполнителей словари прошли редакционно-издательскую обработку (при участии Ю. П. Косарской), в ходе которой были унифицированы их структура и формат. Содержание словарей предусматривает следующие разделы:

- **Введение**, в котором представлено основание разработки, исходные материалы, используемые обозначения и основные параметры словаря;
- **Основные рубрики ГРНТИ** – перчислены подрубрики данного раздела, послужившие основой для лекическогонаполнения словаря;
- **Связанные рубрики других классификаций** – перечислены рубрики других классификаций, связанные с данным разделом ГРНТИ;
- **Указатель основных терминов** – список заглавных терминов словарных статей с указанием страницы словаря и с активной гиперссылкой на неё;
- **Определения основных терминов** – основной раздел, содержащий словарные статьи, расположенные в алфавитном порядке заглавных терминов;
- **Дополнительные ключевые слова** – факультативный раздел, содержащий термины, важные для описания содержания документов данной тематики, но пока не нашедшие авторитетного определения.

Из выше изложенного видно, что, несмотря на единство формата, словари получились не однородными по своему наполнению. Они сильно различаются по объёму: от минимума в 28 статей (науковедение) до максимума в 775 статей (машиностроение). Это различие определяется объёмом общественной практики в данной отрасли, а также степенью детализации знаний в классификационных системах. Общее число статей во всех словарях превосходит 9 тысяч.

Словари во всех случаях обладают следующими общими свойствами:

– содержат термины, отражающие определённую научную область с подробностью, заданной существующими системами индексирования научных работ;

– термины соотнесены с классификационными рубриками, которые являются входами в определённые разделы существующих информационных ресурсов.

Отсюда ясно, что эти словари могут быть использованы не только для содействия в понимании соотношений классификационных рубрик, но и для поиска и навигации по системе разнородных информационных ресурсов, на основе используемой лексике. Для этого необходимо разработанную совокупность словарей совместить с совокупностью таблиц соответствия классификаций в одной сетевой системе. Модель такой Объединённой классификационно-словарной системы (ОКСС) разработана совместно с Н. Н. Шабуровой на материале физики полупроводников[4,5].

Тезаурус тематических рубрик (ТТР) по физике полупроводников и нанотехнологий, разработанный в Институте физики полупроводников СО РАН, содержит наименования ряда классификационных систем и выделенные из них термины. Эти лексические единицы связаны системой тезаурусных ссылок, фиксирующих отношения синонимии, иерархии (родовидовой и партитивной) и ассоциации (по значимому пересечению объёмов понятий). В нём представлены соответствующие тематике рубрики следующих классификаций: УДК, ББК, ГРНТИ (аналогично заданию Минобнауки), а также Рубрикатор ВИНИТИ, PACS (Рубрикатор Американского физического общества) и рубрикаторы Федерального нано-портала. Каждая рубрика представлена своим полным наименованием, а также выделенными из наименований отдельными терминами. Полные наименования снабжены кодами «материнских» классификаций. Для каждой рубрики указаны тезаурусные ссылки «Выше» и «Ниже», в соответствии с их местом в классификационной иерархии. Кроме того в соответствии с экспертным заключением указываются смысловые связи рубрик разных классификаций, аналогично процедуре нынешней работы. Такие же связи указываются и для терминов, выделенных из рубрик. При этом в первую очередь устанавливаются связи между термином и рубрикой, из которой этот термин выделен. Чаще всего рубрика рассматривается как нижестоящий дескриптор по отношению к выделенному из неё термину. Кроме того термину приписывается индексы рубрик, в наименования которых термин входит. Последнее даёт ещё один

[4] Сопоставительный тезаурус классификационных систем по физике полупроводников / В. Н. Белоозеров, Н. Н. Шабурова // Информационное обеспечение науки: новые технологии: Сб. науч. трудов / Н. Е. Калёнов (ред). – М.: Науч. Мир, 2009. – С. 311–322.

[5] Тезаурус тематических рубрик по физике полупроводников как модель объединённой классификационно-словарной системы / В. Н. Белоозеров, Н. Н. Шабурова // 19-й научно-практический семинар «Информационное обеспечение науки: Новые технологии». Таруса, 24–28 августа 2015 г.

канал семантической связи дескрипторов нашего тезауруса. Многие термины снабжены определениями.

Такую же структуру мы предлагаем реализовать на материале словарей и сопоставительных таблиц нынешней работы. Полученные данные полностью соответствуют структуре тезауруса тематических рубрик. Реализовать эту структуру целесообразно в единой базе данных, загрузив туда как словари, так и сопоставительные таблицы. Сейчас, однако, эти материалы загружаются на серверы разных организаций на разном программном обеспечении. Опыт создания базы данных словарей в БЕН РАН представлен на семинаре в Тарусе докладом М. М. Якшина[6]. Там же А. В. Шапкин рассказал о развитии системы классификационных схем в ВИНИТИ РАН, где теперь представлены таблицы соответствия классификаций[7]. Я думаю, что после завершения этих работ следует обменяться результатами, чтобы каждая организация имела полную информацию.

При наличии такой объединённой классификационно-словарной системы (назовём её сокращением ОКСС) её можно использовать и как средство информационной поддержки административных задач в области научно-технической информации, так и как средство лингвистического обеспечения программ навигации по сетевым информационным ресурсам. Независимо от того на каком поисковом языке задан запрос (будь то наименование или код рубрики, или же отдельные термины), связи в ОКСС позволяет выйти на релевантные информационные массивы независимо от того каким языком индексирования систематизированы массивы в информационных ресурсах – той или иной классификацией, ключевыми словами или тезаурусом. Вопрос лишь в том, чтобы в ОКСС были представлены необходимые классификации, термины и их связи. В этом смысле имеющаяся в настоящее время совокупность словарей и таблиц соответствия нуждается в дальнейшем развитии. В достаточной мере сейчас отражены связи классификаций только с ГРНТИ. Можно конечно использовать ГРНТИ как язык посредник для связи между остальными классификациями, но при этом будут складываться неточности соответствий, и было бы лучше разработать таблицы непосредственных соответствий всех классификаций друг с другом. Но главное, что требует большой работы, это установление тезаурусных связей внутри массива отдельных терминов.

Предлагаемое целевое состояние ОКСС видится таким: В единой базе данных представлены элементы трёх основных классов: термины, наименования рубрик, коды рубрик. Между этими элементами установлены обычные тезаурусные связи, отражающие их тематическое наполнение: синонимия (тематическое совпадение), тематическое включение и тематическое пересечение. Кроме того терминам могут быть сопоставлены тексты толкований, а толкованиям сопоставлены библиографические (гипер)ссылки на источники. Кодам и наименованиям рубрик также целесообразно сопоставить ссылки на соответствующие классификационные системы.

Вопрос об использовании этих связей при дальнейшем внедрении в поисковую систему требует отдельного рассмотрения. При этом нужно иметь в виду, что при формальном единстве смысла тезаурусных отношений их программное использование может различаться в зависимости от того, между элементами какого типа та или иная связь установлена. Отношение между термином и классификационной рубрикой в тезаурусе по полупроводникам (установленное на основании вхождения термина в наименование) не тождественно такому же отношению в ныне созданных словарях (которое установлено на основании использования термина для индексирования документов данной рубрики).

Полноценное создание предлагаемой Объединённой классификационно-словарной системы потребует довольно значительных усилий, но уже имеющиеся таблицы соответствий классификаций и привязанные к ним терминологические словари позволяют при умеренных затратах средств и времени реализовать систему тематической навигации по информационным ресурсам с хорошими показателями эффективности.

---

[6] Создание базы данных терминологических словарей / М. М. Якшин // XIX научно-практический семинар «Информационное обеспечение науки: новые технологии». Таруса, 24–28 августа 2015 г.

[7] Разработка предложений по созданию реляционной модели данных взаимосвязанных классификаторов / А. В. Шапкин // XIX научно-практический семинар «Информационное обеспечение науки: новые технологии». Таруса, 24–28 августа 2015 г.