

О некоторых особенностях создания таблиц соответствия между ГРНТИ и УДК

*О. А. Антошкова, Т. С. Астахова, Е. Ю. Дмитриева, О. В. Смирнова
Всероссийский Институт научной и технической информации
Российской Академии Наук (ВИНИТИ РАН),
Москва, Россия*

Ключевые слова: библиографические классификации, иерархические классификации, базовые соответствия, рубрикатор, универсальная десятичная классификация, научно-техническая информация.

Для установления единства и совместимости описания содержания научно-технической информации Минобрнауки поддержало проект ВИНИТИ РАН «Сопоставление ГРНТИ с другими классификационными системами с целью совершенствования системы тематической классификации НИР, НИОКР гражданского назначения. Формирование системы соответствий между различными классификациями в сфере научно-технической информации» (шифр темы 2014-14-573-0024-001, уникальный идентификатор RFMEFI60114X0001).

В рамках проекта разработаны таблицы соответствия ГРНТИ со следующим классификационными системами: УДК, ББК, МПК, ОКСВНК, рубрикаторами ОЭСР, WoS, SCOPUS, РИНЦ, РФФИ, РНФ, РГНФ, ФАНО.

При разработке системы соответствий между различными классификаторами в сфере НТИ базой для сравнения содержательного, структурного, терминологического и кодового соответствия выбран Государственный рубрикатор научно-технической информации.

Государственный рубрикатор научно-технической информации (ГРНТИ) представляет собой универсальную иерархическую классификацию областей знаний, принятую для систематизации потока научно-технической информации. Являясь иерархической классификацией, рубрикатор построен так, что классы одного уровня, как правило, не пересекаются и взаимно исключают друг друга. Понятия одного уровня находятся в состоянии подчинения к понятию более высокого уровня. В рубрикаторе также использовано сочетание иерархии с фасетным принципом, проявляющимся в наличии совокупности рубрик, повторяющихся в разных классах в виде группы «Общие вопросы» и применении типовых классификационных делений в разных разделах Рубрикатора.

Такой выбор определен особенностями ГРНТИ как классификационной системы:

- универсальный охват научного и технического знания, сравнительно неглубокая иерархия (3 уровня),
- пополняемый принцип деления классов, т.е. позволяющий иметь в сети классов резервные места для развития системы,
- согласованность структуры классов со структурой управления народным хозяйством.

В процессе формирования системы соответствий между различными классификациями необходимо было решить следующие задачи:

- установить базовые соответствия между ГРНТИ и другими классификационными системами;
- представить массивы базовых соответствий в виде реляционной базы данных.

Создаваемая в рамках проекта система соответствий между различными классификаторами является, по сути, инструментом многопозиционных взаимосвязей полного или частичного соответствия и представляет собой ряд разработанных сравнительных таблиц смысловых связей между рубриками ГРНТИ и каждой из классификационных систем, входящих в установленный Заказчиком «Перечень основных классификационных систем».

Для установления соответствий между классификационными системами сформирована специальная база данных на SQL-сервере ВИНИТИ РАН, куда загружены массивы сопоставляемых классификаций с той степенью полноты, которая достаточна для указания на соответствие рубрикам ГРНТИ.

Основной целью нашей работы в рамках НИР «Разработка базовых соответствий между ГРНТИ и другими классификационными системами» было выполнение технического задания по установлению соответствий между классификационными системами ГРНТИ и УДК, которые широко применяются в области научно-технической информации.

Универсальная десятичная классификация (УДК) охватывает весь универсум знаний, не являясь при этом конгломератом отдельных отраслевых классификаций. Эта классификационная система успешно применяется для систематизации и поиска разнообразных источников информации в различных по объему и по назначению фондах. УДК, появившаяся в результате дальнейшего развития Десятичной классификации М. Дьюи, сохранила в своей основе иерархическую структуру. Вместе с тем, в процессе развития в УДК возникли дополнения и методические приемы, характерные для фасетной или аналитико-синтетической классификации. Например, Вспомогательные таблицы общих и специальных определителей, позволяющие единообразно строить индексы в соответствии с категориями места, времени, языка или группировать документы в соответствии с отраженными в них технологическими процессами и операциями.

Для этого использовались только актуальные классы УДК, соответствующие современному международному эталону таблиц на английском языке. Поэтому необходимо было сверить классы из ранее опубликованных Таблиц УДК на русском языке (печатная версия) с электронным Эталонным массивом УДК на русском языке, поддерживаемым ВИНТИ РАН и международным англоязычным эталоном Master Reference File, включающим все изменения и дополнения к УДК за последние годы.

В том случае, если определенный класс УДК из печатных таблиц оказывался исключенным из современного эталона, он заменялся актуальным классом соответствующего содержания.

Например, большие изменения произошли в методике применения класса *37 Воспитание. Обучение. Образование* и при присвоении индексов УДК Рубрике ГРНТИ *14 Народное образование. Педагогика* надо было устанавливать соответствия с учетом всех произошедших в УДК изменений. Так, для подрубрики *14.15 Система образования* можно было бы применить класс УДК *371 Организация воспитания и образования*, но этот класс был исключен из таблиц УДК с рекомендуемой заменой на вновь введенный специальный определитель *37.09 Организация обучения*. Этот определитель и был указан в таблице соответствий.

В Таблице для установления более точных соответствий рубрик ГРНТИ индексам УДК приводятся как простые классы из Основных таблиц УДК, так и комбинированные индексы, состоящие из двух и более классов УДК из Основных и Вспомогательных таблиц. В этом случае в графе, предназначенной для текстовой рубрики, относящейся к определенному классу УДК помещено, только условное обозначение «к.и.» (комбинированный индекс), а ключевые слова в данной строке таблицы не приводятся.

Комбинированный индекс раскладывается на составляющие его элементы, которые приводятся в нижних строках, соответствующих рассматриваемой рубрике ГРНТИ, причем к каждому элементу комбинированного индекса даются ключевые слова (в предназначенной для этого графе). Одной рубрике ГРНТИ может соответствовать несколько простых или комбинированных индексов УДК. В этом случае все перечисленные индексы, относящиеся к одной рубрике, отделяются друг от друга точкой с запятой, а ключевые слова к каждому элементу индексов приводятся отдельно в соответствующей графе.

Помимо общей методики применения УДК в данном проекте было целесообразно применить специальные методические приемы, позволяющие в дальнейшем более широко использовать компьютерные технологии для установления классификационных соответствий или анализа классов УДК. По этой методике многие сложные и комбинированные индексы представлены как ряд более простых индексов или отдельных классов УДК отделяемых точкой с запятой или: (знаком отношения).

Например:

Рубрике ГРНТИ *38.59.21 Оборудование горно-разведочных работ* присвоен комбинированный индекс *550.812:550.8.07*, состоящий из классов УДК:

550.812 Поисквые и разведочные работы

550.8.07 Оборудование. Установки и аппаратура

Однако по общей методике применения УДК этот индекс должен представлять собой объединенный класс и быть записан слитно как 550.812.07.

Таким образом, в таблице соответствий многие индексы УДК представляют собой методически адаптированную машиночитаемую версию аналитической росписи сложных (составных) индексов.

Обратная таблица соответствий УДК-ГРНТИ по объему меньше, чем прямая таблица ГРНТИ-УДК. Это обусловлено отсутствием в обратной таблице комбинированных индексов УДК (к.и.). Оставлены только простые индексы, а также общие и специальные определители, обеспечивающие корректные соответствия между рубриками ГРНТИ и классами УДК. Также удалены повторы пары «индекс УДК – рубрика ГРНТИ». Эти повторы присутствуют в прямой таблице, когда одной рубрике ГРНТИ соответствует несколько комбинированных индексов УДК. Комбинированные индексы могут включать один и тот же простой индекс УДК. Например:

76.01.09	История медицины и здравоохранения. Персоналия			x	61(091)	к.и.
76.01.09				x	61	Медицинские науки
76.01.09				x	(091)	Изложение материала строго в историческом аспекте
76.01.09				x	61(092)	к.и.
76.01.09				x	61	Медицинские науки
76.01.09				x	(092)	Изложение материала в виде биографии

Обратная таблица:

61(091)	к.и.	76.01.09	История медицины и здравоохранения. Персоналия
61	Медицинские науки	76.01.09	
(091)	Изложение материала строго в историческом аспекте	76.01.09	
61(092)	к.и.	76.01.09	
61	Медицинские науки	76.01.09	
(092)	Изложение материала в виде биографии	76.01.09	

В данном примере повтор выделен цветом и в дальнейшем удален.

Связь между ГРНТИ и УДК установлена методом интеллектуального анализа практики индексирования и поиска научной информации в БД ВИНТИ РАН и анализом смыслового соответствия рубрик. Установлено четыре вида смысловой связи соответствующих рубрик:

- эквивалентность по содержанию;
- включение объема рубрики ГРНТИ в сопоставленную рубрику;
- включение объема сопоставленной рубрики в рубрику ГРНТИ;
- значительное пересечение объемов рубрик (ассоциация).

Из имеющихся четырех возможных видов соответствий преимущественно был использован вид «ассоциация». Для этого есть объективные основания, которые обусловлены различиями в структуре двух сравниваемых классификаций. ГРНТИ более простая, чем УДК классификационная система, как по структуре, так и по содержательному наполнению. В свою очередь ГРНТИ содержит много актуальных подрубрик, не имеющих в УДК простых прямых аналогов.

Так, например, если по формальному признаку рубрика ГРНТИ 15 Психология и класс УДК 159.9 Психология эквивалентны друг другу, то содержательный анализ данной рубрики и класса

УДК дает нам только ассоциативный вид соответствия. С одной стороны класс УДК 159.9 более развит формально и содержательно, с другой стороны и рубрика ГРНТИ 15 Психология имеет в некоторых своих подрубриках более развитое современное наполнение. Например, подрубрика ГРНТИ 15.81.43 Военная психология может быть обозначена только комбинированным индексом УДК 159.9:355, состоящим из двух простых классов, где класс 355 Военное дело. Подрубрика 15.81.35 Психология управления тоже не имеет простого эквивалента в таблицах УДК, а конструируется с помощью комбинированного индекса из классов 159.9 Психология и 005 Управление. Менеджмент.

Таблицы были подготовлены с учетом необходимой глубины классификации международной статистики по науке и технике. По результатам работы будут выработаны рекомендации для дальнейшего развития тематических классификаций НТИ.

Говоря об общих особенностях экспертного анализа связей ГРНТИ с другими классификациями, следует отметить, что глубина и дробность развития классификаций, входящих в «Перечни», сильно варьирует – от алфавитного перечня нескольких десятков рубрик в классификаторах SCOPUS до практически бесконечной возможности углубления и комбинирования иерархии рубрик в УДК.

Установление соответствий рубрик ГРНТИ рубрикам УДК, ББК, МПК и ОКСВНК проведено параллельно, согласно разным процедурам, что позволяет учесть разнообразные аспекты описания информационных ресурсов по науке и технике.

В ходе выполнения работы разработана специальная методика построения индексов. По этой методике сложные и комбинированные индексы представлены как ряд более простых индексов или отдельных классов УДК, отделяемых точкой с запятой.

Таким образом, в таблице соответствий многие индексы УДК представляют собой методически адаптированную машиночитаемую версию аналитической росписи сложных (составных) индексов.

Обратная таблица соответствий УДК-ГРНТИ построена по аналогичному шаблону, но в нем меняются местами графы рубрики ГРНТИ и УДК, а также строки вышестоящих и нижестоящих соответствий.

Статистика прямой таблицы соответствий ГРНТИ-УДК:

Всего выявлено 19659 соответствий, распределяющихся следующим образом:

№№	Вид соответствия	Код соответствия в таблице	Кол-во индексов
1	Рубрики ГРНТИ и индексы УДК эквивалентны, т.е. относятся к одной и той же теме	=	508
2	Рубрики ГРНТИ выше индексов УДК	>	107
3	Рубрики ГРНТИ ниже индексов УДК	<	80
4	Рубрики ГРНТИ и индексы УДК находятся в ассоциации, т.е. их тематика пересекается существенным образом	x	18964
		Итого	19659

ВЫВОДЫ

Основным принципом построения соответствий является обеспечение синонимичности двух классификаторов через семантическое сопоставление двух ветвей иерархии. Таблицы соответствий обеспечивают взаимодействие различных классификационных систем через систему смысловых соответствий тематических рубрик.

Смысловые связи между классификационными рубриками представлены отношениями логического включения и пересечения объемов понятий и реализованы в виде тезауруса тематических рубрик, что позволит национальным информационным ресурсом оптимизировать управление.

Работы по установлению базовых соответствий между ГРНТИ и другими классификационными системами в сфере научно-технической информации и представлении массива базовых соответствий в виде реляционной базы данных или документов XML проводились строго в соответствии с «Требованиями к работам и их результатам на выполнение научно-исследовательских работ». Алгоритм выполнения работ был определен методикой установления соответствий рубрик ГРНТИ классам основных классификационных систем научной и технической информации на основе смыслового анализа рубрик. В результате, при выполнении научно-исследовательских работ достигнуты следующие результаты:

установлены базовые соответствия между ГРНТИ и другими классификационными системами в сфере научно-технической информации, представленные массивами базовых соответствий в виде документов XML. Полученные результаты уточнены и скорректированы в ходе выполнения работ и совокупной обработки данных.

Формирование соответствий между классификационной системой ГРНТИ и другими классификациями, применяемыми в области научно-технической информации, и выработка рекомендаций для дальнейшего развития тематических классификаций НТИ осуществлялись в целях информационного обмена и повышения результативности сектора научных исследований и перспективных разработок.