АНТОПОЛЬСКИЙ Александр Борисович -
доктор технических наук, профессор,
главный научный сотрудник, Центр по изучению
проблем информатики ИНИОН РАН,
член редколлегии журнала
«Информационные ресурсы России»
Адрес: 117997, г. Москва, Нахимовский проспект, 51/21
e-mail: ale5695@yandex.ru

БЕЛООЗЕРОВ Виктор Николаевич -
кандидат филологических наук, доцент,
ведущий научный сотрудник, ВИНИТИ РАН,
старший научный сотрудник, Вычислительный центр
им. А.А. Дородницына ФИЦ ИУ РАН
Адрес: 125190, г. Москва, ул. Усиевича, 20
e-mail: nomoip@viniti.ru

МАРКАРОВА Тамара Сергеевна -
кандидат филологических наук, доцент
e-mail: tmarkarova@inbox.ru

УДК 002.6:025.48.05

О разработке онтологии на основе классификаторов научной информации и терминологических словарей[1]

Введение

Настоящая работа является продолжением и развитием проекта, выполненного в 2014-2015 гг. по заданию Минобрнауки РФ при головной роли ВИНИТИ[2], в котором принимала участие группа специалистов из разных организаций.

Целью 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» были построение системы соответствий между различными классификаторами в сфере НТИ и выработка рекомендаций для развития системы тематической классификации НТИ в целях оптимизации информационного обмена в научно-технической инновационной сфере и повышения результативности сектора исследований и разработок.

Для достижения поставленной цели были определены и решены следующие задачи:

- разработка методики установления соответствий рубрик ГРНТИ классам основных классификационных систем научной и технической информации на основе смыслового анализа рубрик;
- разработка методики формирования и составление терминологических научных словарей с использованием лексики классификационных систем НТИ;
- установление базовых соответствий между ГРНТИ и другими классификационными системами, представление массива базовых соответствий в виде реляционной базы данных или документа XML;
- составление предложений и рекомендаций по развитию системы тематической классификации научной и технической информации.

Результаты проекта отражены в ряде публикаций [1-4].

О проекте единого российского электронного пространства знаний

В рамках одного проекта по вполне объективным причинам невозможно как охватить весь широкий спектр проблем по систематизации и структуризации научно-информационного пространства, так и в достаточной степени решить поставленные задачи. Поэтому необходимо сформулировать предложения по дальнейшему использованию и развитию полученных результатов с учетом концептуальной постановки задачи, содержащейся в официальных документах.

[1]Работа выполнена при поддержке гранта РФФИ № 17-07-00153 «Исследование системы классификаторов по науке и технике и разработка механизмов смысловой навигации и поиска знаний в информационных сетях».

[2]Шифр НИР «2014-14-573-0024». Руководитель работ - директор ВИНИТИ, академик Ю.М. Арский.

В Указе Президента РФ № 808 от 24.12.2014 «Об утверждении Основ государственной культурной политики» [5] есть такой пункт: «Формирование единого российского электронного пространства знаний на основе оцифрованных книжных, архивных, музейных фондов, собранных в Национальную электронную библиотеку и национальные электронные архивы по различным отраслям знания и сферам творческой деятельности».

Понятие единого российского электронного пространства знаний (далее ЕРЭПЗ) далее перекочевало в изменения закона «О библиотечном деле», принятые в 2016 г. [6], где говорится: «Целью создания Национальной электронной библиотеки являются сохранение исторического, научного и культурного достояния народов Российской Федерации, обеспечение условий для повышения интеллектуального потенциала Российской Федерации и популяризации российской науки и культуры, формирование основы для создания единого российского электронного пространства знаний».

В мае 2017 г. появился проект обновленного Положения о Национальной электронной библиотеке [7], где имеется достаточно развернутая трактовка понятия ЕРЭПЗ. Приведем соответствующий фрагмент:

«III. Порядок формирования единого российского электронного пространства знаний на основе НЭБ.

18. Единое российское электронное пространство знаний (далее - пространство знаний) представляет собой совокупность взаимно интегрированных на основе НЭБ информационных систем и иных информационных ресурсов, сформированных на базе научного, исторического и культурного достояния народов России, а также лучших образцов зарубежных научных и культурно-исторических ценностей, функционирующих на основе единых информационных технологий и принципов, обеспечивающих семантическое связывание их содержимого, а также инструменты поиска и извлечения информации по запросу пользователей.

19. Целью создания пространства знаний является формирование единой, целостной и авторитетной совокупности накопленных человечеством знаний, повышение интеллектуального потенциала Российской Федерации, популяризация российской культуры и науки, в том числе за рубежом. Приоритет при формировании пространства знаний должен отдаваться документам на русском языке и языках народов России.

20. Основными принципами формирования пространства знаний являются:

1) отсутствие ограничения доступа пользователей к информации, содержащейся в пространстве знаний;

2) безвозмездность доступа пользователей к информации;

3) достоверность и авторитетность информации;

4) семантическая взаимосвязанность информации.

21. Пространство знаний формируется на основе следующих компонентов:

1) НЭБ;

2) общенационального научно-образовательного интерактивного энциклопедического портала;

3) информационных ресурсов, содержащих электронные копии документов Архивного фонда Российской Федерации, доступ к которым не ограничен в соответствии с законодательством Российской Федерации;

4) федеральной государственной информационной системы «Государственный каталог Музейного фонда Российской Федерации»;

5) информационных ресурсов, содержащих электронные копии музейных предметов и музейных коллекций музеев Российской Федерации;

6) информационных ресурсов, содержащих электронные копии аудиовизуальных документов, находящихся в ведении организаций, уполномоченных на постоянное хранение аудиовизуальных документов в соответствии с законодательством Российской Федерации;

7) информационной технологии системы классификации, поиска и извлечения информации.

22. Порядок формирования пространства знания включает:

а) создание НЭБ;

б) создание системы классификации, поиска и извлечения информации;

в) взаимную интеграцию на основе НЭБ информационных систем и информационных ресурсов, функционирующих по единым с НЭБ принципам и общим правилам, включающим информационные ресурсы, перечисленные в подпунктах 2-6 пункта 21 настоящего Положения, семантическое связывание содержащихся в них электронных копий документов посредством использования системы классификации, поиска и извлечения информации, а также информационно-телекоммуникационных сетей;

г) создание портала пространства знаний, обеспечивающего извлечение информации по запросам пользователей из входящих в пространство знаний информационных систем и информационных ресурсов - сайта в информационно-телекоммуникационной сети Интернет».

В настоящей статье не ставится задача критического анализа понятия ЕРЭПЗ. Будем исходить из того, что, будучи упомянутым в Федеральном законе, оно тем самым приобрело силу закона. Поэтому следует обсуждать возможные практически и технологически приемлемые трактовки этого понятия, особенно возможные способы реализации *системы классификации, поиска и извлечения информации,* упомянутой в цитируемом выше проекте Положения о НЭБ.

С учетом развития когнитивных информационных технологий в последние годы представляется наиболее вероятным, что *«система классификации, поиска и извлечения информации ЕРЭПЗ»* должна создаваться и поддерживаться с помощью технологий семантической сети и связанных открытых данных, качественное развитие которых в перспективе должно обеспечиваться созданием некоторой онтологии.

С другой стороны, сосредоточение большого научно-информационного массива в таких информационных системах, как Национальная электронная библиотека, другие национальные и отраслевые электронные библиотеки и архивы, является объективным условием для того, чтобы упомянутая онтология стала интегрирующим инструментарием, обеспечивающим взаимодействие с информационно-поисковыми языками, использованными при формировании существующих электронных информационных ресурсов. Причем при создании онтологии необходимо учитывать использование разных классификаций для разнородных ресурсов. Так, для библиотечных ресурсов это, прежде всего, библиотечные классификации - УДК и ББК, для патентных ресурсов - МКИ, для электронных библиотек диссертаций и рефератов - классификация ВАК, для многих ресурсов НТИ - ГРНТИ и т. д.

Таким образом, основу ЕРЭПЗ должна составлять единая русскоязычная онтология научного знания, включающая лексику и парадигматику классификаций, тезаурусов, систем метаданных и других семантических средств, практически используемых для формирования национальных электронных информационных ресурсов - библиотечных, архивных, музейных и других массивов научной и образовательной информации.

Созданная в результате проведенной ранее работы таблица соответствия научных классификаций, снабженная словарями дефиниций, взятых в основном из различных энциклопедий, может, на наш взгляд, послужить исходным материалом для начала работ по созданию онтологии, поскольку эта таблица сохраняет преемственность со многими практически используемыми классификациями.

Онтологии в современной научной литературе

При все нарастающем объеме научной и образовательной информации, при непрерывных качественных и функциональных изменениях этой информации неизбежно встает вопрос о назревшей необходимости принципиально новых, современных способов обработки и структурирования научно-образовательного пространства. В последнее время с учетом экспоненциального роста web-технологий, семантических технологий исследователи все чаще обращаются к онтологиям как к самому адекватному и корректному лингвистическому инструментарию. Существует достаточное количество определений онтологии, иногда противоречивых, иногда мало чем отличающихся от определений тезаурусов. Рамки и тема статьи не позволяют авторам развернуть полноценную дискуссию по поводу определения понятия «онтология» и в конечном итоге высказать свое отдельное мнение и определение. Поэтому мы ограничимся тем, что рассмотрим это понятие подробней, опираясь в основном на обобщающую работу [8]. В частности, в ней говорится:

«В проектировании онтологий условно можно выделить два направления, до некоторого времени развивавшихся отдельно. Первое связано с представлением онтологии как формальной системы, основанной на математически точных аксиомах. Второе направление развивалось в рамках компьютерной лингвистики и когнитивной науки. Там онтология понималась как система абстрактных понятий, существующих только в сознании человека, которая может быть выражена на естественном языке (или средствами какой-то другой системы символов). При этом обычно не делается предположений о точности или непротиворечивости такой системы.

Таким образом, существует два альтернативных подхода к созданию и исследованию онтологий. Первый (формальный) основан на логике (предикатов первого порядка, дескриптивной, модальной и т.п.). Второй (лингвистический) основан на изучении естественного языка (в частности, семантики) и построении онтологий на больших текстовых массивах, так называемых корпусах.

В настоящее время данные подходы тесно взаимодействуют. Идет поиск связей, позволяющих комбинировать соответствующие методы. Поэтому иногда бывает сложно отделить лексические онтологии с элементами формальных аксиоматик от логических систем с включениями лингвистических знаний».

Авторы цитированной работы выделяют три основания классификации онтологий:

• *Классификация по степени формальности*. По этому основанию разделяются системы представления понятий по степени формализации: от строгой формализованной системы, основанной на аксиоматике, до обычного словаря или словника, предназначенного для восприятия человеком.

• *Классификация по цели создания*. В рамках этой классификации выделяют уровни: онтология верхнего уровня, онтология предметной области и прикладная онтология. Они существенно различаются методологией и техникой создания.

• *Классификация по наполнению, содержимому*. По этому основанию онтологии делятся на:

• общие (такие как онтологии верхнего уровня);

• онтологии, ориентированные на предметы;

• онтологии, ориентированные на задачи (функции), в том числе предназначенные для автоматической обработки текста, в частности, лексические онтологии.

Вообще процедура сопоставления понятий и языковых выражений является одной из центральных проблем теории и практики создания онтологий.

В цитированной выше монографии Б.В. Доброва и др. приводится детальный обзор наиболее масштабных проектов онтологий верхнего уровня. Здесь мы их только перечислим с соответствующими ссылками.

OpenCyc - открытая для общего пользования часть коммерческого проекта Cyc, в рамках которого создана наиболее масштабная и детализированная на текущий момент онтология в области здравого смысла. База знаний OpenCyc содержит информацию из различных предметных областей: Философия, Математика, Химия, Биология, Психология, Лингвистика и т. д. Файл с описаниями OpenCyc имеет объем около 700 мегабайт и доступен для скачивания с сайта проекта[3].

DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) - первая из онтологий в библиотеке базовых онтологий проекта WonderWeb[4], которую предполагается применять в Semantic Web для согласования между интеллектуальными агентами, использующими разную терминологию.

SUMO (Standard Upper Merged Ontology) - онтология верхнего уровня, разработанная в рамках проекта IEEE SUO (IEEE Standard Upper Ontology) и Teknowledge[5]. Проект претендует на статус стандарта для онтологий верхнего уровня. Онтология SUMO содержит наиболее общие и самые абстрактные концепты, имеет исчерпывающую иерархию фундаментальных понятий (около 1 тыс.), а также набор аксиом (примерно 4 тыс.), определяющих эти понятия. Назначение SUMO - содействовать улучшению интероперабельности данных, извлечения и поиска информации, автоматического вывода и обработки естественного языка.

Онтология Джона Совы (J. Sowa's ontology), предложенная им в книге «Knowledge Representation: Logical, Philosophical, and Computational Foundations», определяет базовые онтологические категории, полученные автором из источников по логике, лингвистике, философии и искусственному интеллекту[6].

Лингвистический ресурс WordNet[7] разработан в Принстонском университете США. WordNet относится к классу лексических онтологий, свободно доступен в интернете, и на его основе были выполнены тысячи экспериментов в области информационного поиска. WordNet версии 2.1 охватывает приблизительно 155 тысяч различных лексем и словосочетаний, организованных в 117 тысяч понятий или совокупностей. Заметим, что существует и развивается российский аналог этого ресурса[8].

CIDOC CRM («Committee on Documentation» «Conceptual Reference Model»)[9] - онтология в области документации в сфере культурного наследия; представляет собой формальную онтологию, предназначенную для улучшения интеграции и обмена гетерогенной информацией по культурному наследию. Более конкретно, CIDOC CRM определяет семантику схем баз данных и структур документов, используемых в культурном наследии и музейной документации, в терминах формальной онтологии.

Для того чтобы реализовывать различные онтологии, необходимо разработать языки их представления, имеющие исчерпывающую выразительную мощность. Распространение онтологического подхода к представлению знаний оказало содействие при создании разнообразных языков представления онтологии и инструментальных средств, предназначенных для их редактирования и анализа. Существуют традиционные языки спецификации онтологий: *Ontolingua, CycL*, языки, основанные на дескриптивных логиках (такие как *LOOM*), языки, основанные на фреймах *(OKBC, OCML, F-Logic)*. Более поздние языки основаны на Web-стандартах *(XOL, SHOE, UPML)*. Специально для обмена онтологиями через Web были созданы языки *RDF, RDFS, DAML+OIL, OWL*. Именно последний язык получил наибольшее распространение при проектировании онтологий в пространстве Web. *OWL* с 2004 года является рекомендацией W3C и объединяет лучшие черты своих предшественников.

Важным инструментальным средством для работы с онтологиями являются языки запросов к хранилищам онтологий. Наиболее популярными среди языков запросов к RDF-хранилищам на сегодняшний день являются языки *RDQL* и *SPARQL*.

При создании онтологий целесообразно пользоваться инструментальными программными средствами, созданными специально для проектирования, редактирования и анализа онтологий, - редакторами онтологий. Количество редакторов онтологий уже перевалило за 100. В цитируемой выше монографии Б.В. Доброва и его коллег описываются следующие основные редакторы онтологий:

Ontolingua - кроме собственно редактора онтологий, эта система содержит:

• сетевой компонент Webster, предназначенный для определения концептов;

• сервер, обеспечивающий доступ к онтологи-

[3] *http://www.opencyc.com*
[4] *http://www.loa-cnr.it/DOLCE.html*
[5] *http://ontologyportal.org/*
[6] *http://www.jfsowa.com/ontology/*
[7] *http://wordnet.princeton.edu*
[8] *http://www.phil.pu.ru/depts/12/RN/index_ru.shtml*
[9] *http://www.cidoc-crm.org/*

ям Ontolingua по протоколу OKBC (Open Knowledge Base Connectivity);

• Chimaera - инструментарий для анализа и объединения онтологий.

Protege. Это свободно распространяемая Java-программа, предназначенная для построения (создания, редактирования и просмотра) онтологий той или иной прикладной области. Она включает редактор онтологий, позволяющий проектировать онтологии, разворачивая иерархическую структуру абстрактных и конкретных классов и слотов. Данный инструмент поддерживает использование языка OWL и позволяет генерировать HTML-документы, отображающие структуру онтологий.

DOE - (Differential Ontology Editor) - простой редактор, который позволяет пользователю создавать онтологии.

OntoEdit - инструментальное средство, обеспечивающее просмотр, проверку и модификацию онтологии. Оно поддерживает языки представления онтологии OIL и RDFS.

OilEd - автономный графический редактор онтологий, разработанный в рамках проекта On-To-Knowledge. Он свободно распространяется по общедоступной лицензии GPL.

WebOnto представляет собой Java-апплет и разработан для просмотра, создания и редактирования онтологий. Для моделирования онтологий он использует язык OCML (Operational Conceptual Modeling Language).

ODE, WebODE ODE (Ontological Design Environment) взаимодействует с пользователями на концептуальном уровне, обеспечивает их набором таблиц для заполнения (концептов, атрибутов, отношений) и автоматически генерирует код на языках *LOOM, Ontolingua и F-Logic.* Данный инструмент получил свое развитие в редакторе онтологий WebODE.

Среди российских разработок особое место занимает фундаментальное исследование В.Ш. Рубашкина [16]. В нем, кроме обзора и анализа современных подходов к созданию онтологий, ориентированных на поиск и другие задачи автоматической обработки текста, содержится описание языка представления знаний InFol, методики формирования онтологий и онторедактора InTes, разработанных коллективом прикладных лингвистов Санкт-Петербургского государственного университета. Этот коллектив в течение многих лет возглавлял В.Ш. Рубашкин, к великому сожалению, недавно ушедший от нас.

Существенно, что разработки этого коллектива были ориентированы на информационный анализ научных и деловых текстов, а создаваемые онтологии должны опираться на классификации, тезаурусы, другие информационно-поисковые языки и энциклопедические словари. То есть подход В.Ш. Рубашкина и его коллег к созданию онтологий в значительной степени коррелирует с условиями реализации описываемого проекта и мнением его участников.

Система соответствий классификационных систем и рубрикаторов при поддержке логико-понятийных схем терминологических словарей будет особенно ценным источником или компонентом для создания единой русскоязычной онтологии в сфере исследований и разработок.

Создание системы соответствий целесообразно реализовать в формате связанных открытых данных.

Связанные открытые данные (Linked Open Data, LOD) можно определить как связанные наборы данных, опубликованные в RDF-формате и доступные для свободного использования всеми пользователями без каких-либо ограничений в виде авторских прав, патентов и других механизмов контроля.

В 1999 г. консорциум W3C опубликовал набор открытых стандартов Семантической паутины, включающий в себя описание модели RDF. Этот стандарт и стал использоваться в проекте DBpedia, составившем «ядро» пространства наборов данных Linked Open Data, которое также называют LOD-облаком (LOD-cloud). Проект DBpedia стартовал в 2007 г.

В проекте по созданию системы соответствия классификаций, продолжением которого служит данная работа, предусмотрено формирование соответствий рубрик классификационных систем в виде RDF-троек, что позволяет перейти от системы соответствий к форматам LOD без особых трудностей. Таким образом, синтаксис таблиц соответствия будет полностью соответствовать синтаксису LOD.

Что же касается семантики, то лексика классификаторов, объединенная с дефинициями этой лексики за счет энциклопедических и терминологических словарей, должна составить важный фасет единой русскоязычной онтологии в сфере исследований и разработок.

О возможной организации работ в долгосрочной перспективе

Дальнейшая работа в указанном направлении должна предусматривать:

• определение перечня информационных систем и электронных архивов, образующих единое российское электронное пространство знаний и информационных языков, используемых для их создания;

• дополнение таблиц соответствия классификациями, словарями, тезаурусами и системами метаданных, выделенными на предыдущем этапе;

• дополнение полученной БД лексикой и парадигматикой из энциклопедических источников;

• загрузку полученной онтологии в систему представления онтологии и связанных открытых данных (например, в систему СКАН);

• организацию технической поддержки полученной онтологии;

• организацию коллективной работы по развитию и коррекции универсальной онтологии.

Таким образом, основной целью продолжения работ по созданию системы соответствия классификаторов научной информации и терминологических словарей должна быть единая русскоязычная онтология верхнего уровня в сфере исследований и разработок. Эту онтологию целесообразно реализовать и разместить в интернете в формате связанных открытых данных.

Список литературы:

1. Каленов Н.Е., Белоозеров В.Н. Формирование терминологических словарей по лексике классификационных систем // Научно-техническая информация. Сер. 1. Организация и методика информационной работы. - 2015. - № 3. - С. 60 - 70.

2. Арский Ю.М., Никольская И.Ю., Гоннова С.М. Формирование системы тематической классификации с целью развития информационного обмена в научно-технической сфере // Международная конференция Крым-2015 «Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса», г. Судак, 7-13 июня 2015 г.

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

4. Антопольский А.Б., Ефременко Д.В. Инфосфера общественных наук России // Директ-Медиа. - 2017 (в печати).

5. Указ Президента РФ от 24 декабря 2014 г. № 808 «Об утверждении Основ государственной культурной политики» [Электронный ресурс]. - Режим доступа: http://base.garant.ru/70828330/ (дата обращения 28.06.2017).

6. Федеральный закон «О библиотечном деле» от 29.12.1994 № 78-ФЗ ст. 18.1 (ред. от 03.07.2016) [Электронный ресурс]. - Режим доступа: http://fzrf.su/zakon/o-bibliotechnom-dele-78-fz/st-18.1.php (дата обращения 28.06.2017).

7. Проект Постановления Правительства Российской Федерации «Об утверждении Положения о Национальной электронной библиотеке» (подготовлен Минкультуры России 07.08.2016 [Электронный ресурс]. - Режим доступа: http://www.unkniga.ru/images/docs/2017/polozhene-neb-2-variant.pdf (дата обращения 28.06.2017).

8. Онтологии и тезаурусы: модели, инструменты, приложения: учебное пособие / Б.В. Добров, В.В. Иванов, Н.В. Лукашевич, В.Д. Соловьев. - М.: Интернет-университет информационных технологий, Бином. Лаборатория знаний. - 2013. См. также электронную версию http://www.intuit.ru/studies/courses/9/270/inf.

9. Рубашкин В.Ш. Онтологическая семантика. Знания. Онтологии. Онтологически ориентированные методы информационного анализа текста. - М.: Физматлит, 2013. - 348 с.

НАША ИНФОРМАЦИЯ

Премия РУНЕТА

Премия Рунета - старейшая и самая престижная награда в российском сегменте сети Интернет - отмечает в этом году свое 14-летие. Церемония награждения лауреатов пройдет в ноябре 2017.

Премия Рунета является общенациональной наградой в области высоких технологий и интернета, поощряющей выдающиеся заслуги компаний-лидеров в области информационных технологий и электронных коммуникаций, государственных и общественных организаций, бизнес-структур, а также отдельных деятелей, внесших значительный вклад в развитие российского сегмента сети Интернет (Рунета).

Основные номинации:

1. «Технологии и инновации»
2. «Экономика, бизнес и инвестиции»
3. «Культура, СМИ и массовые коммуникации»
4. «Государство и общество»
5. «Здоровье и досуг»
6. «Наука и образование».

Специальные номинации:

1. «Экология и окружающая среда»
2. «Народное голосование».

Гран-при HOT-LIST 2018

Помимо лауреатов в основных и специальных номинациях, в рамках Премии Рунета в этом году впервые будут определены трендсеттеры (организаторы и проекты) в 10 технологических, бизнесовых и инновационных направлениях - они войдут в HOT-LIST 2018 по версии Премии Рунета.

Подробная информация о Премии Рунета-2017 доступна на официальном сайте: *http://premiaruneta.ru*