Об одном подходе к формированию предметных онтологий различных областей науки

Н.Е. Каленов[1]

[1] *Межведомственный суперкомпьютерный центр РАН - филиал ФГУ ФНЦ Научно-исследовательский институт системных исследований (НИИСИ) РАН*

Аннотация. Под предметной онтологией в контексте статьи понимается совокупность ключевых понятий, относящихся к некоторой области науки, с их семантическими связями (тезаурус), дополненная индексами различных классификационных систем, связанными с ней. Предметные онтологии являются необходимой составляющей каждого подпространства, входящего в Единое цифровое пространство научных знаний (ЕЦПНЗ). Составление и поддержка предметной онтологии любой научной области - огромная работа, осложняющаяся динамичностью современной науки. Во многих областях не только естественных, но и общественных наук постоянно возникают новые термины и связи между ними. Эти изменения опосредованно отражаются в научных публикациях. В этой связи задача поддержки предметных онтологий требует постоянного мониторинга новых научных публикаций, скорость нарастания количества которых по всем областям знаний растет год от года. В данной статье предлагается методология выделения новых ключевых терминов отдельной области науки, базирующаяся на использовании существующих классификационных систем в совокупности с базами данных (БДЦ) цитирования, такими как WEB of Science и Scopus для англоязычных публикаций и Российский индекс цитирования (РИНЦ) – для русскоязычных. Методология предполагает разбиение научной области на ряд разделов в соответствии с выбранной классификационной системой, выделение из БДЦ ядра статей, относящихся к каждому разделу, а из статей – новых авторских ключевых терминов, которые и должны составлять, в совокупности с соответствующими разделами классификационных систем, основу предметной онтологии данной научной области.

Ключевые слова: пространство научных знаний, предметная онтология, базы данных цитирования, ключевые термины, тезаурус, классификационные системы

About one approach to the formation of subject ontologies for science various fields

N.E. Kalenov[1]

[1] *Joint Supercomputer Centre of Russian Academy of Sciences (JSCC RAS)*

Abstract. The subject ontology (in the context of the article) is understood as a combination of key concepts related to a certain field of science, with their semantic connections (thesaurus), supplemented by indices of various classification systems associated with this scientific field. Subject ontologies are a necessary component of each subspace that is part of the Common Digital Space of Scientific Knowledge (CDSSK). Compiling and supporting the subject ontology of any scientific field is a huge job. This work is complicated by the fact that new terms and relations between them are constantly emerging in many areas of not only the natural sciences, but also in the social sciences. These changes are being reflected in scientific publications. In this regard, the task of supporting subject ontologies requires constant monitoring of new scientific publications in order to identify new scientific terms. This article proposes a methodology for highlighting new keywords in a particular field of science, based on the use of existing classification systems in conjunction with citation databases(CSB) such as WEB of Science and Scopus (for English keywords) and Russian Citation Index (for Russian keywords). The methodology involves dividing the scientific field into a number of sections in accordance with the selected classification system, extracting from the CSB the core of articles related to each section, and from the articles - new author's keywords, which should constitute, in combination with the corresponding sections of classification systems, the basis of the subject ontologies of this scientific field.

Keywords: scientific digital space, subject ontology, citation databases, keywords, thesaurus, classification systems

Одним из важных направлений современной информатики, связанных с сохранением и распространением достижений науки, является создание единого цифрового пространства научных знаний (ЕЦПНЗ). В этом пространстве должны быть отражены достоверные знания, полученные в различных областях науки; цель создания пространства – предоставление пользователям различных категорий многоаспектной информации как внутри отдельных научных направлений, так и на стыке наук. В соответствии с концепцией, отраженной в [1-3], ЕЦПНЗ представляет собой совокупность разнородных информационных ресурсов, сгруппированных в тематические подпространства, объединяемые единой онтологией. Под единой онтологией понимаются общие для всех подпространств принципы их построения - единые

подходы к хранению и предоставлению информации, формированию классов объектов и связей между ними, профилей метаданных и атрибутов, пользовательских интерфейсов и т.д. Онтология и поддерживающая ее программная оболочка ЕЦПНЗ должны обеспечивать развитый многоаспектный поиск разнородной информации, удобную ее визуализацию и навигацию по связанным ресурсам. Основой для тематического поиска информации в каждом подпространстве должна служить его предметная онтология – набор связанных терминов, описывающих данную область науки. Проблема формирования предметных онтологий тесно связана с традиционными задачами формирования тематических тезаурусов. Этим задачам посвящено достаточно много исследований – как зарубежных, так и отечественных [4 - 6]. Разработаны стандартные формы представления тезаурусов в машинном виде [7], программные средства для их формирования и встраивания в электронные библиотеки [8]. Предметная онтология любого подпространства ЕЦПНЗ должна содержать тезаурус в качестве необходимой составляющей. Но, поскольку в ЕЦПНЗ предполагается включать разнородные ресурсы, в том числе, извлекаемые из существующих баз данных и каталогов библиотек, в состав предметной онтологии должны входить индексы различных классификационных систем, описывающие ресурсы данного научного направления, таких, как ГРНТИ, УДК, МКИ и др.

В 2017-2019 гг, при поддержке РФФИ (проект 17-03-12013-ОГН) был создан прототип предметной онтологии по ряду научных направлений [9-12], реализованный в виде действующей системы терминологических словарей [13]. В основу онтологий были положены термины и их определения, соответствующие понятиям, отраженным в ГРНТИ [14]. По нескольким научным направлениям между терминами были установлены дефинитивные связи, и система была дополнена индексами УДК и ББК.

Работы в этом направлении были продолжены при поддержке РФФИ (проект 18-00-00372). Для развития прототипа предметной онтологии до полноценной версии необходимо дополнить ее ключевыми терминами (КТ), достаточно детально описывающими выбранные научные направления. Эту задачу предлагается решать, используя ключевые термины, которые указывают авторы в своих научных статьях. Для автоматизации этого процесса и отбора качественных публикаций могут быть использованы базы данных цитирования, такие как WEB of Science (WoS) и Scopus (для англоязычных публикаций) и Российский индекс цитирования (РИНЦ) – для русскоязычных.

Предлагаемая методика включает следующие процессы.

1. Разбиение области науки на отдельные разделы. Степень детализации такого разбиения определяется учеными совместно с информационными работниками на основе анализа существующих

классификационных систем (КС), таких как ГРНТИ, УДК, МКИ и др., достаточно хорошо проработанных для данной научной области.

2. Установление связей между выделенными разделами и относящимися к данной области науки индексами выбранных КС. По отношению к каждому разделу в каждой КС выбираются индексы, в той или иной мере связанные с данным разделом. Формируются пары «раздел – индекс» и внутри пар устанавливаются связи одного из 4-х типов: тождественны, содержит, содержится, пересекаются.

3. По каждому из выделенных разделов формулируются запросы к БД цитирования (WoS, Scopus, РИНЦ), в соответствии с которыми выбираются публикации за определенный интервал годов, зависящий от научного направления. Метаданные каждой статьи, полученной по запросу, имеют атрибуты, содержащие перечень ключевых терминов и ссылки на публикации, цитирующие данную статью.

Полученные данные могут служить основой для формирования предметной онтологии. При этом возможны различные подходы к отбору ключевых терминов, подлежащих включению в нее. Можно выделить и проанализировать все ключевые термины, полученные на этом этапе; определить частотность встречаемости каждого и выбрать наиболее часто встречаемые. Второй вариант предусматривает фильтрацию публикаций путем выделения наиболее цитированных в качестве источников ключевых терминов.

Технически отбор ключевых терминов из WoS и Scopus не представляет сложности – обе эти системы могут обрабатывать запросы в автоматическом режиме предоставляют возможность получения информации в различных структурированных форматах, позволяющих легко выделять авторские ключевые термины и количество цитирующих работ.

Несколько сложнее обстоит дело с РИНЦ, в котором не предусмотрены ни автоматическая обработка запросов, ни выдача информации в каком-либо структурированном формате. Система выдает данные в виде текстовых записей. Для извлечения необходимой информации потребуетсяразработка программы, обрабатывающей HTML-страницы, содержащие найденные публикации.

В результате обработки полученной информации по каждому выделенному разделу данной научной области формируется массив русскоязычных (РИНЦ) и англоязычных терминов (WoS, Scopus) с указанием частоты их встречаемости в течение любого заданного интервала времени.

В первом приближении можно считать, что все выделенные ключевые термины входят в соответствующие разделы данного научного направления, которые, в свою очередь, связаны установленным ранее видом связи с индексами различных КС. Очевидно, что полученные

наборы ключевых терминов требуют редактирования со стороны специалистов в данном научном направлении, но эта работа существенно проще, чем поиск ключевых терминов.

Предлагаемая технология позволяет не только формировать предметные онтологии, но и может служить основой для наукометрического анализа развития отдельных направлений той или иной области науки. На основании сопоставительного анализа набора и частотности ключевых терминов, полученных в текущем и предыдущих годах, можно делать выводы о динамике развития тех или иных областей науки.

Предлагаемая методика была опробована в 2019 году на примере моделирования предметной онтологии по микробиологии [15]. В этом научном направлении было выделено 42 раздела. По каждому из них на основании обработки статей, полученных по запросам к БД WoS, выбраны ключевые термины. Их распределение по разделам микробиологии представлено в табл. 1.

Название раздела	**Число отобранных статей**	**Число КТ в разделе**
Общие вопросы микробиологии	136	748
Методы и аппаратура в микробиологии	77	586
Систематика и номенклатура микроорганизмов	70	1181
Морфология, цитология, физиология и биохимия микроорганизмов. Общие вопросы.	108	711
Морфология, цитология и циклы развития микроорганизмов; иммунохимия микробной клетки	78	732
Физиология микроорганизмов	200	645
Биохимические процессы микроорганизмов	116	461
Коммуникативные межклеточные взаимодействия у микроорганизмов	176	672
Рост и культивирование микроорганизмов	150	1592
Действие внешних факторов на микроорганизмы	95	641
Генетика и селекция микроорганизмов. Методы исследований	188	1411

Генетика бактерий	230	1336
Генетика дрожжей и микроскопических грибов	402	323
Экология микроорганизмов	250	593
Водная микробиология	197	614
Почвенная микробиология	250	740
Геомикробиология	248	664
Роль микроорганизмов в очистке окружающей среды	119	549
Симбиоз и антагонизм у микроорганизмов. Взаимоотношения микрорганизмов с насекомыми, беспозвоночными и др.	34	75
Биология возбудителей заболеваний человека и животных. Общие проблемы	166	893
Биология бактерий-возбудителей заболеваний человека и животных	145	618
Биология грибов-возбудителей заболеваний человека и животных	54	213
Неклассифицированные малоизученные микроорганизмы	169	510
Взаимоотношение возбудителя и хозяина	235	484
Клиническая микробиология	51	564
Лекарственная чувствительность микроорганизмов	150	614
Лабораторная диагностика бактериальных инфекций и микозов	173	418
Микробиология внутрибольничных, раневых и других инфекций. Терапия и профилактика	50	548
Техническая микробиология. Общие проблемы	78	516
Оборудование для микробиологических производств. Системы контроля и управления промышленными процессами микробного синтеза	87	78
Промышленное получение биологически активных веществ микробиологическим путем	120	119

Биодеградация, биоконверсия и ферментация	108	513
Микробиология пищевых продуктов	150	327
Микробная деградация технических материалов, загрязняющих веществ и других химических веществ	80	372
Биогеотехнология	29	92
Медицинские проблемы микробиологических производств	51	194
Сельскохозяйственная микробиология	113	328
Космическая биологии	109	124
Микробиологическая очистка окружающей среды	126	89
Санитарная микробиология	126	307
Бактерийные препараты	117	149
Эпидемиология микроорганизмов	254	371

Табл. 1 Распределение ключевых терминов по разделам микробиологии.

Всего было обработано 5865 статей, из которых выделено 22715 различных ключевых англоязычных терминов. Ближайшая перспектива – профессиональный перевод этих терминов на русский язык и объединение всех КТ с рубриками классификационных систем по микробиологии в единой базе данных.

Полученные результаты базируются на исследованиях, поддерживаемых РФФИ (проекты №№ 18-00-00372 и 20-07-00103).

Литература

1. Антопольский А.Б., Каленов Н.Е., Серебряков В.А., Сотников А.Н. О едином цифровом пространстве научных знаний // Вестник Российской академии наук. 2019. Т. 89. № 7. С. 728-735.
2. Антопольский А.Б. и др. Принципы построения и структура единого цифрового пространства научных знаний (ЕЦПНЗ) // Научно-техническая информация. Сер. 1. 2020. № 4. С. 9 – 17.
3. Kalenov, N., Sobolevskaya, I., Sotnikov, A. Mathematical modeling of the processes of interdisciplinary collections formation in the digital libraries environment // CEUR Workshop Proceedings. 2020. pp. 391-398.
4. M. Mercedes Martínez-González, María Luisa Alvite Díez. The support of constructs in thesaurus tools from a Semantic Web perspective: Framework

to assess standard conformance // Comput. Stand. Interfaces, 2019 Iss. 65: Pp. 79-91
5. C. Roche, R. Costa, S. Carvalho, B. Almeida. Knowledge-based terminological e-dictionaries: The EndoTerm and al-Andalus Pottery projects. // Terminology, International Journal of Theoretical and Applied Issues in Specialized Communication, 2019. № 2. Pp 259–290.
6. Белоозеров В.Н. Гуревич И.Б., Трусова Ю.О. Тезаурус по анализу изображений в сети терминологических словарей // Перспективные направления исследований и критические технологии в классификационных системах : материалы конф. Москва, 2017 С. 35-36.
7. UNESCO Thesaurus https://skos.um.es/unescothes/ (дата обращения 28.04.2020)
8. О. М. Атаева, В. А. Серебряков. Персональная открытая семантическая цифровая библиотека LibMeta. Конструирование контента. Интеграция с источниками LOD // Информатика и её применения, 2017, том 11, выпуск 2, страницы 85–100
9. Антопольский А.Б., Белоозеров В.Н., Маркарова Т.С. О разработке онтологии на основе классификаторов научной информации и терминологических словарей // Информационные ресурсы России. 2017. № 5 (159). С. 2-7.
10. Antopolskiy A.B. and others. The Development of a Semantic Network of Keywords Based on Definitive Relationships // Scientific and Technical Information Processing. - 2017, Vol.44, No.4, pp.261-265.
11. Антопольский А.Б., Белоозеров В.Н., Каленов Н.Е., Маркарова Т.С. О развитии терминологической базы данных в виде комплекса отраслевых информационно-поисковых тезаурусов // Информационные ресурсы России, 2018. - № 5 (165). - С. 22-30.
12. Белоозеров В.Н., Шабурова Н.Н. О разработке классификационно-тезаурусной онтологии для предметной области физики и радиоэлектроники // Информационное обеспечение науки: новые технологии: сб. науч. тр. Екатеринбург, 2018,. С. 75-86
13. N.E. Kalenov, A.M. Senko. Interactive system of terminological dictionaries as one of the elements in the ontology of scientific knowledge. // Software Journal: Theory and Applications (electronic Journal), 2019. Iss. 4. http://swsys-web.ru/en/interactive-system-of-terminological-dictionaries.html (дата обращения 28.04.2020)
14. Государственный рубрикатор научно-технической информации. http://grnti.ru (дата обращения 28.04.2020).
15. Цветкова В.А., Харыбина Т.Н., Мохначева Ю.В., Бескаравайная Е.В., Митрошина И.Ю. Особенности совмещения классификационных систем и формирования массива ключевых слов для определения пространства знаний по микробиологии // Научные и технические библиотеки, 2019. - № 11. - С. 25-43.

References

1. Antopolskij A.B., Kalenov N.E., Serebryakov V.A., Sotnikov A.N. O edinom cifrovom prostranstve nauchnyh znanij // Vestnik Rossijskoj akademii nauk. 2019. Iss. 89. № 7. pp. 728-735.
2. Antopolskij A.B. and others. Principy postroeniya i struktura edinogo cifrovogo prostranstva nauchnyh znanij (ECPNZ) // Nauchno-tehnicheskaya informaciya. Ser. 1. 2020. № 4. pp. 9 – 17.
3. Kalenov, N., Sobolevskaya, I., Sotnikov, A. Mathematical modeling of the processes of interdisciplinary collections formation in the digital libraries environment // CEUR Workshop Proceedings. 2020. pp. 391-398.
4. M. Mercedes Martínez-González, María Luisa Alvite Díez. The support of constructs in thesaurus tools from a Semantic Web perspective: Framework to assess standard conformance. Comput. Stand. Interfaces, 2019 Iss. 65: pp. 79-91
5. C. Roche, R. Costa, S. Carvalho, B. Almeida. Knowledge-based terminological e-dictionaries: The EndoTerm and al-Andalus Pottery projects Terminology // International Journal of Theoretical and Applied Issues in Specialized Communication, 2019. № 2. pp. 259–290.
6. Beloozerov V.N. Gurevich I.B., Trusova Yu.O. Tezaurus po analizu izobrazhenij v seti terminologicheskih slovarej // Perspektivnye napravleniya issledovanij i kriticheskie tehnologii v klassifikacionnyh sistemah : materialy konf. Moskva, 2017 pp. 35-36.
7. UNESCO Thesaurus https://skos.um.es/unescothes/ (28.04.2020)
8. O. M. Ataeva, V. A. Serebryakov. Personalnaya otkrytaya semanticheskaya cifrovaya biblioteka LibMeta. Konstruirovanie kontenta. Integraciya s istochnikami LOD // Informatika i eyo primeneniya, 2017, tom 11, vypusk 2, pp. 85–100
9. Antopolskij A.B., Beloozerov V.N., Markarova T.S. O razrabotke ontologii na osnove klassifikatorov nauchnoj informacii i terminologicheskih slovarej // Informacionnye resursy Rossii. 2017. № 5 (159). pp. 2-7.
10. Antopolskiy A.B. and others. The Development of a Semantic Network of Keywords Based on Definitive Relationships // Scientific and Technical Information Processing. - 2017, Vol.44, No.4, pp.261-265.
11. Antopolskij A.B., Beloozerov V.N., Kalenov N.E., Markarova T.S. O razvitii terminologicheskoj bazy dannyh v vide kompleksa otraslevyh informacionno-poiskovyh tezaurusov // Informacionnye resursy Rossii, 2018. - № 5 (165). - pp. 22-30.
12. Beloozerov V.N., Shaburova N.N. O razrabotke klassifikacionno-tezaurusnoj ontologii dlya predmetnoj oblasti fiziki i radioelektroniki // Informacionnoe obespechenie nauki: novye tehnologii: sb. nauch. tr. Ekaterinburg, 2018,. pp. 75-86

13. N.E. Kalenov, A.M. Senko. Interactive system of terminological dictionaries as one of the elements in the ontology of scientific knowledge. // Software Journal: Theory and Applications (electronic Journal), 2019. Iss. 4. http://swsys-web.ru/en/interactive-system-of-terminological-dictionaries.html (28.04.2020)
14. Gosudarstvennyj rubrikator nauchno-tehnicheskoj informacii. http://grnti.ru (28.04.2020).
15. Cvetkova V.A., Harybina T.N., Mohnacheva Yu.V., Beskaravajnaya E.V., Mitroshina I.Yu. Osobennosti sovmesheniya klassifikacionnyh sistem i formirovaniya massiva klyuchevyh slov dlya opredeleniya prostranstva znanij po mikrobiologii // Nauchnye i tehnicheskie biblioteki, 2019. - № 11. - pp. 25-43.