ИНФОРМАЦИОННО-ПОИСКОВЫЕ СИСТЕМЫ

УДК 025.44/.47+026.06

О. А. Лаврёнова, В. В. Павлов
РГБ

Библиотечно-библиографическая классификация как традиционная система организации знаний в среде открытых связанных данных

Задачи представленного в статье проекта – опубликование классификационной модели знаний в виде связанных открытых данных (Linked Open Data, LOD), а также обеспечение доступа к ней непосредственно из пространства Semantic Web стандартными средствами сети. В качестве источника данных для классификационной модели знаний принята отредактированная система разделителей систематического каталога РГБ, работающего на основе Библиотечно-библиографической классификации. Рассмотрены разработанная структура данных на базе модели SKOS, а также выбранное программное обеспечение. На конкретных примерах продемонстрированы принципы работы модели с выходом на поиск в каталогах электронной библиотеки и традиционного фонда РГБ.

Ключевые слова: библиотечно-библиографическая классификация, Semantic Web, связанные открытые данные, LOD, система навигации, информационный поиск, модель организации знаний, электронный каталог, электронная библиотека, SKOS.

UDC 025.44/.47+026.06

Olga Lavrenova and Vasili Pavlov
Russian State Library, Moscow, Russia

Library Bibliographic Classification as a traditional knowledge organization system in the linked open data environment

The task of the project is to introduce classification knowledge model as Linked Open Data, LOD, and to provide access to it from Semantic Web through standard network instruments. The reviewed system of RSL systematic catalog (Library Bibliographic Classification) subject divisions is accepted as a source for classification knowledge model The authors examine SKOS-based data structure and the software. Operation principles in terms of the search in e-library catalog and RSL traditional collections are discussed.

Keywords: Library Bibliographic Classification, Semantic Web, linked open data, LOD, information retrieval, knowledge organization model, electronic catalog, e-library, SKOS.

The knowledge organization systems support a number of important features of digital libraries: (1) displaying the content and indexing; (2) information search on the basis of the structured knowledge; (3) provision of a roadmap for semantic knowledge branches and disciplines; (4) provision of tools for interaction of conceptual schemes; (5) the conceptual keystone for the knowledge-based systems. To ensure the interoperability of such resources in the global network environment there are used knowledge organization systems (Linked Data vocabularies). To enrich user queries on search resources our library has began to publish the classification model of knowledge in the form of linked open data with access to them directly from the Semantic Web space and the integration with other classification vocabularies. We took the National Library Bibliographic Classification Code (LBC) as the basics for enrichment the search query with the additional features that are reflected in the electronic catalog of the Russian state library. At the same time this technology performs search in the content of digital library. Full description of LBC indices is introduced into bibliographic record, thereof e all words of hierarchical tree are used in processing of query. We have developed the navigator for separators of the library general systematic catalog. The navigation system allows you to view the hierarchy of the classification sections of both the upper and any other levels, in the words of the verbal language index (in any grammatical form), as well as directly in the indices. Navigator can be seen as a virtual directory systematically associated with the database of electronic catalog and electronic library. When searching for an arbitrary combination of query words, the navigation gives the chain of verbal formulations. The user marks necessary topics and receive information about the number of documents for each of them. It is important that the presence or absence of "decoding" the index in bibliographic records do not matter. Navigator selects those in which the request coincides with the required index, or at least with its initial signs.

Возможности классификаций
при поиске электронных ресурсов

Сетевые системы организации знаний (*Networked Knowledge Organization Systems, NKOS* [**1**]) в последнее десятилетие играют всё более серьёзную роль в управлении информацией в электронной форме. К ним относят классификационные системы, тезаурусы, лексические базы данных, онтоло-

гии и таксономии. Первые два типа систем нашли наиболее широкое применение, так как их проще создать. Однако во всём библиотечном мире удалось сделать универсальными только несколько классификаций.

Перечисленные системы организации знаний обеспечивают целый ряд важных функций электронных библиотек (ЭБ) [**2**]:

представление содержания и индексирование информации и документов,

поддержка пользователей при поиске информации на основе структурирования знаний,

выполнение роли семантической дорожной карты для областей знаний и дисциплин,

обеспечение средств взаимодействия при создании концептуальных схем,

роль концептуального базиса для систем, основанных на знаниях (особенно классификации).

Для обеспечения интероперабельности таких ресурсов в глобальной сети используются технологии ***публикации систем организации знаний в виде словарей связанных данных*** (*Linked Data vocabularies*). Необходимы оптимальные практические решения относительно методов представления таких словарей и установления связей между ними, а также эффективные сетевые средства и приложения для их реализации.

В условиях полнотекстового поиска документов стали актуальными вопросы: могут ли традиционные методы организации знаний улучшить полнотекстовый поиск, следует ли считать классификацию, категоризацию, приписывание ключевых слов или меток и полнотекстовый поиск взаимодополняющими?

На эти вопросы можно дать положительные ответы, продемонстрировав, как системы организации знаний повышают качество поиска в полнотекстовых базах данных:

обеспечивают обогащение запросов пользователей, направляемых в полнотекстовые БД электронных библиотек, новыми полезными поисковыми признаками, что увеличивает показатели полноты поиска;

создают дополнительные сервисы для пользователя.

Только при наличии обоснованной пользы для поиска имеет смысл тратить силы и средства на создание связанных открытых данных (*Linked Open Data, LOD*) [**3**, **4**] конкретного вида.

Задача рассматриваемого проекта заключается в том, чтобы с целью обогащения запросов пользователей на поиск ресурсов РГБ опубликовать классификационную модель знаний в виде связанных открытых данных. Требуется обеспечить доступ к ним непосредственно из пространства *Semantic Web* (семантической паутины, семантического веба) [**4**–**6**] и использовать возможности интеграции классификации с другими словарями. В первую очередь необходимо выбрать наиболее функциональные способы представления этих данных в семантической паутине.

Работы проводятся при поддержке Российского фонда фундаментальных исследований. Принципы работы в среде *LOD* и результаты первого этапа проекта были представлены в работе [**3**]. В этой статье обоснованы проектные решения и освещены новые результаты 2016 г.

Изначально в качестве фундамента для этих работ была выбрана технология использования элементов национальной Библиотечно-библиографической классификации РФ (ББК) для обогащения запросов дополнительными поисковыми признаками в электронном каталоге РГБ. Одновременно технология поддерживает поиск в её электронной библиотеке. В библиографические записи вносятся полные расшифровки индексов ББК, что делает поисковыми все слова, использованные в соответствующей ветви иерархического дерева классификации по умолчанию (подробно об этой технологии см. в статьях [**4**, **6**, **7**]).

Пример 1:

Территории коллективной идентичности в современном французском дискурсе : автореф. дис. ... д-ра ист. наук : 07.00.07. – Москва, 2010.

Для данного ресурса в ЭБ построен и внесён в библиографические метаданные индекс ББК: *Т52(47Фр)*. Индекс расшифрован в БЗ, т.е. для него построена цепочка словесных формулировок индекса с верхнего уровня до нижнего:

История. Исторические науки -- Этнография -- Этнография современных народов -- Народы Европы -- Народы Южной и Юго-Западной Европы -- Французы. Франция

Автореферат из примера 1 может быть найден, в частности, на запрос «*Этнография народов Южной Европы*». Таким образом и происходит обогащение запроса пользователя для обнаружения документов по темам, более узким по смыслу, чем заданная.

Что касается создания дополнительных сервисов на основе классификационных систем, то обеспечение в ЭБ навигации по иерархии областей знаний (тематических разделов) классификаций с наибольшей очевидно-

стью демонстрирует важную роль классификационных моделей для поиска ресурсов. Результаты такого рода технологий не могут быть компенсированы поиском по полному тексту.

В РГБ разработан действующий проект ***навигатора по разделителям генерального систематического каталога*** (ГСК). Система навигации позволяет просматривать иерархию разделов классификации как с верхнего, так и с любого другого уровня, который будет найден по словам из словесных формулировок индексов (в любой грамматической форме), а также непосредственно по индексам. Навигатор можно рассматривать в качестве некоторого виртуального систематического каталога (СК), связанного с базами данных ЭК и ЭБ.

При поиске по произвольному сочетанию слов запроса навигатор находит цепочки словесных формулировок. Пользователь отмечает интересующие его темы, получает информацию о количестве документов для каждой из них, поднимается или спускается по иерархии, окончательно выбирает тему, и система передаёт индексы ББК в ЭК, где отыскиваются библиографические записи. Важно то, что наличие или отсутствие «расшифровки» индекса в библиографических записях не имеет значения. Навигатор отбирает те из них, в которых индекс запроса совпадает с искомым или хотя бы с его начальными знаками.

Пример 2: на рис. 1 продемонстрирован фрагмент иерархии разделителей виртуального СК выведенного на экран результата поиска по сочетанию слов «*движение, спутник, планеты*».

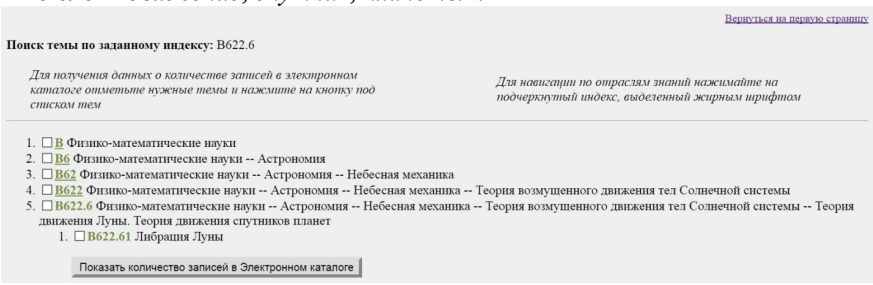

Рис. 1. Вывод на экран результата поиска в навигаторе по словам «*движение, спутник, планеты*»

Особенно интересные результаты получаются при обнаружении тех документов, которым индексы ББК были приписаны до 1998 г. Ранее эти индексы не расшифровывались в библиографических записях ЭК (см. БЗ на автореферат диссертации 1971 г. издания в примере 3).

Пример 3:

Сравнительный анализ развития структурных форм и речных долин Южного Сихотэ-Алиня [Текст] : автореф. дис. на соискание учёной степ. канд. геолого-минерал. наук. – Москва : 1971.

Запись содержит три индекса ББК без словесных формулировок. Вначале навигатор сравнивает индексы, найденные по словам в запросе, с индексами в библиографической БД ЭК и ЭБ, находит, в частности, индексы для автореферата из примера 3. При этом в записи два индекса из трёх совпадают полностью с индексами в навигаторе, а в одном индексе совпадает начальная часть. В примере подчёркнуты найденные части индексов, т.е. автореферат был обнаружен по индексам этих уровней.

Д9(2р55,31)823 *-- Науки о Земле -- Региональный раздел наук о Земле -- Россия, СССР -- Азиатская часть -- Дальний Восток. Дальневосточный экономический район – Приамурье и Приморье -- Сихотэ-Алинь* (частичная расшифровка индекса)
Д823.121.51 *-- Науки о Земле -- Географические науки -- Физическая география -- Геоморфология -- Геоморфология суши -- Экзогенные процессы и обусловленный ими рельеф -- Водно-эрозионный и водно-аккумулятивный рельеф (флювиальный рельеф) -- Речные долины. Эрозионные каньоны и ущелья -- Речные долины* (полная расшифровка всех уровней)
Д392.3 *-- Науки о Земле -- Геологические науки -- Тектоника -- Тектонические движения земной коры -- Складчатые и орогенические движения* (полная расшифровка всех уровней)

Постановка задачи

Нередко поднимается вопрос: стоит ли тратить огромные усилия на преобразование классификаций в современные модели знаний? Существует, по крайней мере, два аргумента в пользу этой работы:

недопустимо потерять средства поиска миллионов документов, содержание которых в библиотеках описано с помощью этих классификаций и которые относятся к культурному наследию страны;

для создания новых структур представления больших массивов ресурсов, универсальных по содержанию, потребуются огромные затраты, участие широкого круга специалистов; за более полувека развития информационных технологий сделать это никому не удалось;

существует общий библиотечный принцип: недопустимо закрытие карточного каталога, если все его поисковые возможности не обеспечены в ЭК.

Опыт РГБ позволил принять в рамках проекта решение о построении классификационной модели в среде связанных открытых данных не на основе эталонных таблиц ББК, а с использованием иерархической структуры оцифрованных разделителей Генерального систематического каталога (ГСК). Следует отметить, что другие известные универсальные классификации (в частности, УДК) публикуются в семантической паутине в исходном виде со всеми примечаниями и ссылками.

Известно, что одно из основных свойств модели любого объекта – соответствие конкретной задаче. При решении иной задачи для отображения свойств объекта потребуется другая модель. Задача описываемой модели – поддержка поиска текстов по классификационным индексам, которые уже сформированы при содержательной обработке этих документов.

Требовалось выявить такие типы элементов классификационных данных, которые повысят качество поиска при условии их присоединения к исходному запросу программным путём, как это реализуется в среде *LOD*. Например, полезно обработать и использовать примечания, приведённые в эталонах таблиц и содержащие уточняющие примеры, а также слова или словосочетания, эквивалентные по смыслу терминам конечного элемента цепочки формулировок индекса. В дальнейшем потребуется обработка данных из оцифрованного алфавитно-предметного указателя к ГСК.

Такой подход обеспечивает преемственность традиционной и автоматизированной технологий, т.е. сохранение и совершенствование в электронной среде тех возможностей тематического поиска книг, диссертаций, картографических и нотных изданий, которые были гарантированы в традиционном систематическом каталоге.

Сделаем основополагающие выводы относительно представления классификационных данных в среде связанных открытых данных:

основа моделирования классификации в семантическом пространстве – установление связи классификационного индекса, построенного при обработке конкретного документа, с его полной «расшифровкой» (цепочкой словесных формулировок индекса) и связей между индексами;

данные для построения классификационной модели в среде *LOD* – это отредактированная система разделителей ГСК РГБ (более 130 тыс. рубрик);

критерий дополнительного включения конкретного типа данных – возможность повышения качества поиска при их присоединении к исходному запросу программным путём.

Результаты анализа ресурсов, представляемых в среде LOD, реализуются в приведённых далее проектных решениях.

Публикация классификации в *Semantic Web*

Для полноты картины вспомним, в чём заключается технология формирования связанных открытых данных [**3, 4**]. Связи в среде семантической паутины устанавливаются между ресурсами. Ресурсами считаются любые данные: текст статьи, БЗ, слово или словосочетание, фраза (например, определение термина), код, индекс, дескриптор, отношение, свойство и т.д. Каждый ресурс получает URI (*Uniform Resource Identifier* – универсальный идентификатор ресурса в сети).

Основу структуры *Semantic Web*, как известно, составляет модель описания данных RDF (*Resource Description Framework*) [**8**], которая позволяет описывать предметную область в терминах ресурсов, свойств ресурсов и значений свойств.

Любое утверждение (*statement*) о ресурсе структурируется, соответственно, в форме триплета (тройки) «*субъект – предикат – объект*». Поскольку субъект, предикат и объект, будучи ресурсами, получают URI, триплеты выглядят как последовательности таких адресов в сети. Так как идентификаторы уникальны, ими может оперировать любой разработчик ресурсов или систем в *Semantic Web* в целом, что и обеспечивает их интероперабельность при установлении связей между ресурсами в сети.

Таким образом, публикация систем организации знаний, в частности классификаций, в LOD реализуется на основе модели данных RDF, базирующейся на языке разметки текстов XML. Уникальные идентификаторы (имена) для обозначения ресурсов выбираются в специальных абстрактных хранилищах (множествах, моделях). Такого рода логически организованное хранилище называется «пространство имён» (*namespace*) [**9**]. Если не удаётся найти в сети логически или семантически подходящее имя, можно создать собственное пространство имён.

Классификации и тезаурусы в RDF

Особенности представления классификаций в RDF хорошо видны при сравнении их с информационно-поисковыми тезаурусами (ИПТ) [**10**]. Дескрипторы тезаурусов или индексы классификаций в своей совокупности обычно рассматриваются как системы координат документов (ресурсов) в условном семантическом пространстве – пространстве знаний. Известно, что в тезаурусах отображаются исключительно парадигматические (не зависящие от контекста) семантические отношения (связи).

Иерархическая структура классификации как система координат в семантическом пространстве принципиально отличается от структуры тезауруса. Координатами документа (ресурса) служат сформированные для него индексы, и они совершенно необязательно являются точными индексами из

классификационной таблицы. Действительной системой координат, состоящей из индексов как имён классов и используемой для размещения документов в семантическом пространстве, предлагаем считать систему готовых индексов, которая на практике формируется в систематических каталогах библиотек.

Поскольку в классификационной модели смысловое содержание индекса выражает полная иерархия его словесных формулировок, в RDF-представлениях для семантической паутины индексу нельзя ставить в соответствие только словесную формулировку нижнего уровня иерархии, как это делается в других проектах представления классификаций в среде LOD, но обязательно ставить полную цепочку словесных формулировок этого индекса. Рассмотрим пример 4, в котором некоторому документу приписан сложный индекс с использованием различных вспомогательных таблиц. Цепочка словесных формулировок индекса демонстрирует, что он отображает не только парадигматические отношения между его составляющими, как это принято в тезаурусах, но и синтагматические (контекстуальные).

Пример 4:

Ш141.2-032я721
Филологические науки. Художественная литература -- Языкознание -- Индоевропейские языки -- Славянские языки -- Восточнославянские языки -- Русский язык -- История языка -- Историческая лексикология -- Этимология -- Учебные пособия для средней школы
«Филологические науки. Художественная литература -- Языкознание» – парадигматические связи между словесными формулировками в цепочке, расшифровывающей индекс.
«Индоевропейские языки -- Славянские языки -- Восточнославянские языки -- Русский язык» -- парадигматические связи.
«История языка -- Историческая лексикология -- Этимология» – парадигматические связи, которые в ИПТ стали бы продолжением первой последовательности.

Однако между выделенными фрагментами в словесной формулировке индекса, а также между ними и фрагментом *Учебные пособия для средней школы* существуют синтагматические отношения.

Индексы кодируют высказывания различной степени сложности по правилам конкретной классификации. Можно провести аналогию с предложением на естественном языке, смысловое содержание которого не равно

простой сумме семантики отдельных слов, его составляющих. Каждый раздел классификации организуется в соответствии со структурой области знания, принятой в конкретной науке, определённой классификации или определённых странах.

Итак, класс, к которому отнесён документ, обозначается полным индексом, приписанным документу при обработке. Словесные формулировки индексов автоматически связывают их со словами естественного языка, на котором пользователь выражает свои запросы. Обогащение запроса происходит на основе использования иерархических и ассоциативных смысловых связей между индексами классификации, учёта грамматических форм, а также отношений синонимии слов, если предметный (словесный) вход в классификацию проработан в этом направлении. Например, требуется добавить слово *лингвистика* для слова *языкознание;* формы *солнцем, солнца, солнцу* и т.д. – для слова *солнце;* прилагательное *солнечный* для существительного *солнце* [**6**].

На первом этапе проекта [**3**] файлы классификации, полученные в результате преобразования отредактированных разделителей СК РГБ в RDF с использованием элементов SKOS, были успешно загружены в семантическое хранилище для последующего манипулирования данными с помощью языка запросов SPARQL. Задача была реализована на базе программного обеспечения *Virtuoso Universal Server* (http://virtuoso.openlinksw.com/).

В международной практике известные классификационные системы представляются в форме, понятной для машины, именно с использованием пространства имён SKOS. Собственно, модель SKOS как приложение RDF и создана «для отображения базовой структуры и содержания таких концептуальных моделей (структур), как тезаурусы, классификационные системы, списки предметных рубрик, таксономии, фолксономии и другие контролируемые словари такого рода» [**10**]. Основным элементом словаря SKOS считается *concept* (концепт) [Там же]. Под концептами подразумеваются единицы смысла – идеи, значения или категории объектов и событий. SKOS различает две основные категории свойств: иерархические (*выше – ниже*) и ассоциативные, которые обозначаются как *related* (связанные). Свойства *выше и ниже* используются для построения прямых иерархических связей между двумя концептами SKOS. Это позволяет программным приложениям удобным способом прокладывать путь от любого концепта до нижестоящего или вышестоящего.

На втором этапе проекта объектом в RDF-описаниях данных используется целый ряд дополнительных свойств. В результате получился следующий состав элементов данных для файлов ГСК, кодируемых в RDF:

URI – skos:Concept

индекс ББК – skos:notation

полная цепочка формулировок индекса – skos:prefLabel

альтернативная цепочка формулировок индекса – skos:altLabel

вышестоящий индекс – skos:broader

нижестоящий индекс – skos:narrower (формируется автоматически)

ссылки «смотри также» и «смотри» – skos:related

примечание, уточняющее содержание индекса и содержащее примеры (более узкие или равнозначные темы или понятия по отношению к выраженному в словесной формулировке данного индекса), – skos:example

последний элемент цепочки формулировок индекса – skos:hiddenLabel (вычленяется программно и копируется из полной цепочки формулировок)

формальные (служебные) элементы для ведения БД (на будущее):

skos:historyNote – описывает существенные изменения смысла или формы концепта

skos:changeNote – документирует структурные изменения относительно концепта (перенос в другое дерево и т.д.).

Следует подчеркнуть, что в ходе исследований принята иная структура значения свойства *skos:prefLabel,* чем используемая в RDF-представлениях ресурсов других классификаций. Объектом такого рода утверждения становится полная цепочка словесных формулировок индекса, а не формулировка нижнего уровня (см. обоснование выше и анализ примера 4). Так, в примере 4 могло быть следующее представление словесной формулировки индекса *Ш141.2-032я721*с помощью свойства *skos:prefLabel:*

<skos:prefLabel xml:lang="ru"> Учебные пособия для средней школы </ skos:prefLabel>.

Новый вариант представления данного индекса*:*

<skos:prefLabel xml:lang="ru"> Филологические науки. Художественная литература -- Языкознание -- Индоевропейские языки -- Славянские языки -- Восточнославянские языки -- Русский язык -- История языка -- Историческая лексикология -- Этимология -- Учебные пособия для средней школы</skos:prefLabel>

Пример 5 демонстрирует предварительное представление в RDF индекса ББК *В253.31/32* (не указываются URI вышестоящих и нижестоящих концептов и некоторые другие данные)*.

Пример 5:

@prefix skos: <http://www.w3.org/2004/02/skos/core#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
<skos:Concept
*rdf:about="*http://lod.rsl.ru/bbkgsk/concepts/%D0%92253.31%2F32*"*> – URI концепта, который обозначается индексом *В253.31/32*
<skos:notation> В253.31/32</skos:notation>
<skos:prefLabel xml:lang="ru"> Физико-математические науки -- Механика -- Механика деформируемых сред -- Гидромеханика и аэромеханика (механика жидких и газообразных сред) -- Гидродинамика и аэродинамика -- Гидродинамика (динамика несжимаемой жидкости)</skos:prefLabel>
<skos:broader xml:lang="ru"> В253.3 </skos:broader>
<skos:narrower xml:lang="ru"> В253.31я4 </skos:narrower>
<skos:narrower xml:lang="ru"> В253.310 </skos:narrower>
<skos:narrower xml:lang="ru"> В253.311 </skos:narrower>
и другие нижестоящие (формируются программно на основе их связи с индексом *В253.31/32* как вышестоящим).
<skos:hiddenLabel xml:lang="ru"> Гидродинамика (динамика несжимаемой жидкости) </skos:hiddenLabel>
<skos:related> В251.3 </skos:related>
</skos:Concept>

Программное обеспечение проекта

Схема классификационных данных, разработанная на втором этапе проекта, а также новые требования, предъявляемые к поиску, выявили необходимость смены программного обеспечения. Предыдущий пакет программного обеспечения *Virtuoso Universal Server* выполнял только следующие функции:

---

* Пояснения к обозначениям:
префиксы *@prefix* указывают сетевые адреса пространств имён, из которых взяты имена свойств концепта;
метки элементов данных заключены в угловые скобки;
сочетание знаков </ обозначает начало конечной метки элемента;
пример триплета: *В253.31/32 (*субъект) – *skos:broader – имеет вышестоящий* (свойство) – *В253.3 (*объект*).*

хранилище семантического представления данных;

программный интерфейс к хранилищу на базе протокола SPARQL.

Учитывая текущие и будущие потребности проекта, были сформулированы дополнительные требования к программному обеспечению. Поиск должен производиться с учётом морфологии русского языка, синонимов.

Запросы с использованием SPARQL позволяют оперировать только регулярными выражениями, поэтому для удовлетворения новых требований к поиску необходимо было найти программное обеспечение, которое бы расширяло протокол SPARQL.

В результате поиска был найден фактически единственный программный пакет, удовлетворяющий всем требованиям, – *Apache Jena*. Он представляет собой платформу для создания приложений связанных данных и семантической паутины. Ниже перечислены компоненты платформы, которые были использованы для реализации сервиса предоставления открытых данных ББК:

высокопроизводительное хранилище семантических данных – *TDB*;

сервис индексации и полнотекстового поиска на базе *Apache Lucene*;

сервер SPARQL – *Fuseki*.

Помимо протокола SPARQL, сервер *Fuseki* поддерживает полнотекстовые запросы (*Jena text query*) к встроенному серверу *Lucene*.

Для демонстрации возможностей обогащения поисковых запросов пользователей создан прототип поисковой системы. В прототипе реализован поиск индексов:

по полным словесным формулировкам;

по примечаниям, уточняющим содержание индекса и содержащим примеры, которые должны быть также поисковыми.

Поиск производится с учётом синонимов и морфологии русского языка. Полученные в результате поиска индексы могут быть использованы в следующих сценариях:

для демонстрации пользователям поисковых подсказок как во время ввода поискового запроса, так и в поисковой выдаче;

для автоматического поиска различных вариантов слов и добавления в результаты поиска документов, содержащих соответствующие индексы.

Например, пользователя интересуют исследования и области применения инфракрасного излучения, включая методы и технические средства, и при вводе слова *инфракрасный* в поисковую строку ему выводится подсказка. В примере 6 представлена такого рода поисковая подсказка с вариантами возможного дополнения запроса индексами ББК, найденными системой по заданному слову (приводятся только некоторые темы из выведенного списка).

Пример 6:

- ***В349*** *Физико-математические науки -- Физика -- Оптика --* ***Инфракрасные*** *лучи*
- ***В652.421.2*** *Физико-математические науки -- Астрономия -- Солнечная система -- Солнце -- Физика Солнца -- Излучение Солнца -- Излучение света Солнцем. Солнечная радиация --* ***Инфракрасное*** *и ультрафиолетовое излучение Солнца*
- ***Г512.311*** *Биологические науки -- Общая биология -- Общая физиология, общая биофизика и общая биохимия -- Общая биофизика -- Биологическое действие физических факторов -- Действие света. Фотобиология -- Действие ультрафиолетовых и* ***инфракрасных*** *лучей*
- ***386-530.12*** *Техника. Технические науки -- Энергетика. Радиоэлектроника -- Радиоэлектроника -- Квантовая радиотехника -- Квантовые приборы -- Лазеры (квантовые генераторы и усилители оптического диапазона) -- Лазеры по диапазону излучения -- Лазеры* ***инфракрасного*** *диапазона*

Другой вариант – предоставление пользователю поисковой подсказки в выдаче одновременно с библиографическими данными. Для запроса *затмение солнца* в поисковую выдачу будет выведено сообщение, приведённое в примере 7. Пользователю предоставляется возможность выбрать для поиска темы, связанные с основной отношением «*смотри также*».

Пример 7:

Возможно, вы ищете документы на тему:

- ***В652.475.3*** *Физико-математические науки -- Астрономия -- Солнечная система -- Солнце -- Физика Солнца -- Атмосфера Солнца -- Солнечная корона -- Исследование короны вне солнечных затмений*
- ***В652.6*** *Физико-математические науки -- Астрономия -- Солнечная система -- Солнце -- Солнечные затмения*
- ***В654.137-3*** *Физико-математические науки -- Астрономия -- Солнечная система -- Планеты и спутники. Планетная астрономия -- Отдельные планеты и спутники -- Планеты типа Земля -- Земля-планета -- Луна -- Лунные затмения и покрытия звёзд Луною*

Выбор связанных ресурсов

Публикацией систем организации знаний в среде связанных открытых данных и выбором принципов установления связей между словарями занимаются многие ведущие библиотеки различных стран. Интересно, что в библиотечном сообществе наиболее туманное представление – о выборе тех ресурсов, с которыми, в принципе, целесообразно и разумно устанавливать связи. Этот вопрос особенно важен, так как создание связанных открытых данных – дорогое и трудоёмкое дело. Относительно классификаций, кроме идеи связывания между собой такого рода словарей разных библиотек и на разных языках, ничего достаточно убедительного не обнаружено.

Что касается выбора ресурсов (словарей), с которыми имеет смысл связывать классификационные данные в среде LOD в рамках рассматриваемого проекта, то предполагается ориентироваться только на такие, которые могут способствовать существенному обогащению изначальных запросов пользователя на поиск в библиографических и полнотекстовых ресурсах.

Для поиска библиографических записей и полных текстов с использованием данных СК по словам и их словосочетаниям, вводимым пользователем, считаем полезными связи с нормативными/авторитетными файлами библиотек для дополнения:

введённого в запрос географического названия другими его вариантами (прежними, новыми, сокращёнными и т.д.);

имени лица другими вариантами (псевдонимов писателей – их подлинными именами, другими псевдонимами и т.д.);

наименования организации другими вариантами (сокращёнными, принятыми официально, прежними, последующими и т.д.).

Найденный системой по запросу индекс из ГСК (полные научные таблицы ББК) целесообразно дополнить индексами среднего варианта и сокращённых таблиц ББК, которые также должны быть представлены в LOD. Это может обеспечить использование классификационной модели для поиска в библиотеках, применяющих соответствующие таблицы. Разумным представляется также установить связь с индексами УДК в RDF [**12**] для обнаружения в семантической паутине (по требованию пользователя) ресурсов, которые снабжены индексами УДК и относятся к нужной области знания. Эти задачи являются предметом дальнейших разработок.

СПИСОК ИСТОЧНИКОВ

1. **Networked** Knowledge Organization Systems. – Режим доступа: http://nkos.slis.kent.edu

2. **Workshop** NKOS – 15th European Networked Knowledge Organization Systems. – Режим доступа: www.tpdl2016.org/nkos

3. **Шварцман М. Е., Найдин О. П.** Linked Open Data как средство обогащения поисковых запросов // Унив. кн. – 2015. – № 12. – С. 66–71.
Shvartsman M. E., Naydin O. P. *Linked Open Data kak sredstvo obogashcheniya poiskovyh zaprosov // Univ. kn. – 2015. – № 12. – S. 66–71.*

4. **Лаврёнова О. А.** Технологии открытых связанных данных и «дорожные карты» как навигаторы для пользователей библиотек // Информ. ресурсы – футурол. аспект: планы, прогнозы, перспективы : материалы X Всерос. науч.-практ. конф. «Электрон. ресурсы библиотек, музеев, архивов», 30–31 окт., 2014 г., Санкт-Петербург. – С.-Петербург : Политехника-сервис, 2014. – С. 146–154.
Lavrenova O. A. *Tehnologii otkrytyh svyazannyh dannyh i «dorozhnye karty» kak navigatory dlya polzovateley bibliotek // Inform. resursy – futurol. aspekt: plany, prognozy, perspektivy : materialy X Vseros. nauch.-prakt. konf. «Elektron. resursy bibliotek, muzeev, arhivov», 30–31 okt., 2014 g., Sankt-Peterburg. – S.-Peterburg : Politehnika-servis, 2014. – S. 146–154.*

5. **Семантическая** паутина (**Semanticheskaya** pautina). – Режим доступа: https://ru.wikipedia.org/wiki/Семантическая_паутина

6. **Лаврёнова О. А.** Возможности пользователя при поиске в электронных библиотеках, или «Витязь на распутье» // Библиотековедение. – 2013. – № 3. – С. 43–52.
Lavrenova O. A. *Vozmozhnosti polzovatelya pri poiske v elektronnyh bibliotekah, ili «Vityaz na raspute» // Bibliotekovedenie. – 2013. – № 3. – S. 43–52.*

7. **Лаврёнова О. А.** Семантические средства библиографического поиска в Российской государственной библиотеке // Общетеорет. и футурол. проблемы библиогр. Библиогр. запись как основа формирования библиогр. ресурсов : материалы II Междунар. библиогр. конгр. «Библиография: взгляд в будущее» (Москва, 6–8 окт. 2015 г.) / Рос. гос. б-ка. – Москва : Пашков дом, 2016. – С. 309–323.
Lavrenova O. A. *Semanticheskie sredstva bibliograficheskogo poiska v Rossiyskoy gosudarstvennoy biblioteke // Obshcheteoret. i futurol. problemy bibliogr. Bibliogr. zapis kak osnova formirovaniya bibliogr. resursov : materialy II Mezhdunar. bibliogr. kongr. «Bibliografiya: vzglyad v budushchee» (Moskva, 6–8 okt. 2015 g.) / Ros. gos. b-ka. – Moskva : Pashkov dom, 2016. – S. 309–323.*

8. **RDF 1.1** Concepts and Abstract Syntax. – Режим доступа: https://www.w3.org/TR/rdf11-concepts/

9. **Namespaces** in XML 1.0 (Third Edition). – Режим доступа: https://www.w3.org/TR/xml-names/

10. **SKOS.** Simple Knowledge Organization System Primer. W3C Working Group Note 18 August 2009. – Режим доступа: http://www.w3.org/TR/skos-primer/

11. **RDF** Schema 1.1. W3C Recommendation 25 February 2014. – Режим доступа: https://www.w3.org/TR/rdf-schema/

12. **УДК (UDC)**. – Режим доступа: http://www.xml.com/pub/a/2005/06/22/skos.html

---

Olga Lavrenova, *Cand. Sc. (Philology), Senior Researcher, Department Head, Russian State Library;*
olavr2009@yandex.ru; lavr@rsl.ru
3/5, Vozdvizhenka st., 119019 Moscow, Russia

Vasili Pavlov, *head of Internet-technologies Support Department, Russian State Library;*
pavlovVV@rsl.ru
3/5, Vozdvizhenka st., 119019 Moscow, Russia