Классификационная модель знаний для обогащения запросов пользователей электронных библиотек

Knowledge classification model to expand e-library users requests

О. А. Лаврёнова
Российская государственная библиотека,
Москва, Россия

Olga Lavreneva
Russian State Library,
Moscow, Russia

Рассматриваются способы обогащения (enrichment) запросов пользователей электронных библиотек на основе формирования классификационных метаданных в системе навигации, имитирующей работу библиотечного систематического каталога, а также предлагается решение задачи перевода этой технологии в среду открытых связанных данных (LOD, Linked Open Data). Элементы классификационной модели представлены в RDF и загружены в семантическое хранилище. Технология поиска связана с электронной библиотекой РГБ в экспериментальном режиме.

The ways to enrich e-library user requests based on classification metadata in navigation imitating library systematic catalog are examined; the solution for transferring this technology into the Linked Open Data (LOD) environment is offered. Classification data elements are encoded and loaded into semantic repository. The technology is applied to the Russian State Library's e-library as a pilot project.

1. Проблемы тематического поиска

1.1. Рассмотрим следующие вопросы относительно тематического поиска в электронных библиотеках (ЭБ):

– На что рассчитывает современный пользователь ЭБ при тематическом поиске текстов публикаций?

– Какие возможности тематического поиска в ЭБ он хочет получить в результате развития технологий ЭБ?

Эти вопросы могут показаться тривиальными, но, на самом деле, ответы на них должны определять развитие автоматизированных информационных систем (АИС).

Обогащение запроса происходит за счёт присоединения к заданным словам их грамматических вариантов, а затем – дополнения найденного классификационного индекса нижестоящими индексами. От последнего режима пользователь имеет возможность отказаться и получить только верхний из найденных индексов.

1.2. Пользователь прежде всего полагается на полноту выдачи документов по заданной в запросе теме.

При работе в Яндексе или Гугле пользователь не рассчитывает на гарантированную полноту результатов поиска, хотя на него обрушивается несколько миллионов сайтов в виде результата его изысканий. Обращаясь в ЭБ серьёзной библиотеки, пользователь полагает, что создатели систем и библиотекари – это добросовестные и квалифицированные специалисты, которые подготовили условия для получения ими достаточно полного результата. Вписывая в поисковую строку слова и их словосочетания, якобы отражающие его информационную потребность, он подспудно надеется на высокие показатели полноты результатов поиска. Точность поиска легко определяется при анализе результатов. Получил 100 документов, а представляются соответствующими запросу 70. Значит, с точки зрения человека, точность поиска – 70%, а информационный шум – 30%. Полноту поиска пользователь на практике не может установить, так как для этого нужно проверить на соответствие данному запросу документы из всей базы данных библиотеки или хотя бы из статистически обоснованной выборки, что тоже для него немыслимо.

Пользователю в голову не придёт, что совершенно не обязательно по запросу «глаголы тюркских языков» он получит в ЭБ публикации по глаголам татарского, якутского, чувашского, тувинского, хакасского, киргизского, камасинского, башкирского, азербайджанского и т.д. языков, поскольку иерархические связи в системе не предусмотрены. Странно было бы ожидать от обыч-

ного человека, что он станет вносить в запрос все тюркские языки через логический знак ИЛИ. В результате он не обнаружит в выдаче публикации о большинстве языков и, что самое печальное, может не догадаться о размере поисковых потерь. Многие думают, что такая проблема касается только поиска по библиографическим записям электронных каталогов, но не в полнотекстовых базах данных. Однако вряд ли кто-то возьмётся со всей уверенностью утверждать, что авторы всех документов по указанной в запросе теме обязательно приведут в текстах всю иерархию вышестоящих терминов или разделов. В результате, такого рода поисковую систему и основанную на ней ЭБ нельзя признать надёжной с точки зрения полноты обеспечения пользователей информацией по заданной теме или предмету. При этом неизвестны случаи, когда библиотеки предупреждали бы об этом читателей.

Проблема решается с помощью, так называемого обогащения (enrichment) запросов на базе различных систем (моделей) организации знаний (KOS, Knowledge Organization Systems): тезаурусов, классификаций, таксономий, онтологий.

1.3. Современный пользователь привык к работе в открытом пространстве Интернет и поэтому ожидает, что в скором времени поиск в электронных каталогах и электронных библиотеках также будет осуществляться в аналогичной среде.

Он считает слишком обременительным выбор конкретных библиотек с различными ЭК и ЭБ, с необходимостью изучения структуры и правил каждой информационной системы. В результате труды могут оказаться напрасными. Спасением могут быть сводные каталоги, но пока их возможности поиска по темам трудно назвать удовлетворительными.

В нашем проекте предлагается одно из возможных решений данной проблемы на основе формирования классификационных метаданных в среде открытых связанных данных (LOD, Linked Open Data) в Семантической паутине (Семантическом вебе, Semantic Web) .

2. Средства обогащения тематических запросов пользователей

2.1. Наиболее мощным методом моделирования знаний являются онтологии, но их разработка представляет собой сложнейшую задачу, решаемую в наше время только для отдельных областей знаний, в основном, хорошо структурированных. Тезаурусы и классификации иногда называют простыми онтологиями. Информационно-поисковые тезаурусы широко распространены в информационных системах, но достаточно развитый универсальный тезаурус до сих пор никому не удалось создать. Представляется, что одна из основных тому причин – отсутствие организационной основы и сотрудничества специалистов различных областей знаний. Библиотекари как наиболее серьёзные и опытные специалисты в области формирования метаданных смогли создать несколько универсальных классификационных систем, которым уделяется большое внимание в сфере использования моделей организации знаний для обогащения запросов.

2.2. Для наглядности предлагается рассмотреть технологию обогащения поисковых запросов на примере одного документа (автореферата диссертации), произвольно выбранного в ЭБ РГБ. Единственное требование при выборе – рассмотреть именно публикацию научного содержания, так как при анализе популярных текстов значение классификационной модели кажется не столь убедительным.

3. Проектные решения РГБ для обогащения запросов

3.1. Пользуясь одним из имеющихся в ЭБ РГБ средств тематического поиска, находим полный текст автореферата диссертации, который и послужит нам примером для анализа принципов автоматического обогащения запросов. Библиографическая запись (БЗ) приводится в сокращённом виде.

Изучение механизма сопряжения синтеза АТФ и протонного транспорта АТФ-синтазой из бактерии RHODOBACTER CAPSULATUS : автореферат дис. ... кандидата биологических наук : 03.00.04 / МГУ . – Москва, 1998 . – 21 с.

ББК: Е472.311.5,0

Е072.511.271-31,0

http://dlib.rsl.ru/rsl01000000000/rsl01000277000/rsl01000277109/rsl01000277109.pdf

Запрос формулировался следующим образом: *«обмен веществ микроорганизмов»*.

Понятно, что данный документ не может быть найденным напрямую по имеющимся в БЗ элементам, которые дают некую информацию о его смысловом содержании. В заглавии нет ни одного слова из запроса, а индекса ББК представляют собой некие коды (индексы), которые расшифровываются только в классификационных таблицах, причём нередко по частям (основные деления, специальные типовые деления, типовые деления общего применения и т.д.). Многие возразят, что в ЭБ автореферат может быть найден по словам из полного текста. В принципе, это возможно, но не в данном случае. Проверено, что автор рассматриваемого научного труда использует в тексте только более узкие термины, непосредственно описывающие его исследования.

Мало что даёт название специальности ВАК:

03.00.04 Биохимия (Биологические, химические, технические, сельскохозяйственные, медицинские, ветеринарные)

Следует отметить, что данный автореферат не будет найден пользователям и по другим вышестоящим элементам иерархии тем, закодированных в индексе *Е472.311.5,0:*

Е472.311.5 Биологические науки -- Микробиология -- Физиология, биофизика и биохимия микроорганизмов -- Биохимия микроорганизмов -- Обмен веществ и энергии у микроорганизмов. Питание микроорганизмов -- Анаболизм (ассимиляция) -- Биосинтез -- Фотосинтез. Фотосинтезирующие микроорганизмы

Расшифровка второго индекса ББК показывает иерархическое описание автореферата в другом аспекте. Понятно, что и по этим темам он без классификационного индекса найден не будет.

Е072.511.27 Биологические науки -- Общая биология -- Общая физиология, общая биофизика и общая биохимия -- Общая биохимия -- Химический состав и химические превращения отдельных веществ живых организмов. Обмен веществ -- Органические вещества -- Азотсодержащие органические соединения. Азотистый обмен -- Нуклеиновые кислоты. Нуклеиновый обмен -- Предшественники и продукты распада нуклеиновых кислот

Приведённые «расшифровки» индексов представляют собой полные иерархические цепочки их словесных формулировок. В тексте автореферата почти все слова из цепочек словесных формулировок индексов отсутствуют, так как автор не посчитал целесообразным показать своим читателям структуру отрасли науки, к которой относится его диссертация.

Проектное решение РГБ

Полные цепочки словесных формулировок индексов ББК вносятся в БЗ с самого начала внедрения ЭК [1]. Все слова в словесных формулировках индексов и их произвольные сочетания являются в ЭК и ЭБ РГБ поисковыми. В результате расшифровки индексов непосредственно в БЗ иерархические связи между темами работают при поиске автоматически (по умолчанию), так что пользователь (читатель) может не задумываться о способе получения результата. Технологически создаётся впечатление, что поиск идёт просто по заданным ключевым словам. Таким образом, обеспечивается иерархический поиск для всех документов, метаданные которых содержат словесные формулировки индексов ББК, т.е. обогащение запроса, автоматически осуществляется за счёт включения в поиск слов из цепочек всех более низких уровней иерархии в индексе ББК.

Слова из словесных формулировок работают при поиске в полнотекстовой базе данных ЭБ РГБ (http://search.rsl.ru) **наравне с другими метаданными.**

3.2. Изначально была поставлена и задача формирования классификационной модели организации знаний на основе ББК [1] поскольку достаточное количество пользователей желают **самостоятельно осуществлять навигацию по иерархическим структурам классификаций, т.е. областей знаний,** и выбирать требуемые темы. Кроме того, не следует забывать те документы, для которых индексы в БЗ не были расшифрованы. Их тоже требуется найти.

Откроем секрет. Автореферат, выбранный в качестве примера, не был найден по цепочке словесных формулировок, поскольку по какой-то причине её не ввели в БЗ. Поэтому предполагаемый пользователь о нём бы не узнал, если бы не специально разработанная система- навигатор, которая имитирует усовершенствованную технологию поиска по систематическому каталогу на основе классификационной системы (модели) представления знаний.

Проектные решения РГБ

Основа классификационной модели – оцифрованная и отредактированная система разделителей Генерального систематического каталога (ГСК) РГБ (порядка 130 000 разделов), а не собственно эталон таблиц для научных библиотек. Работа именно с полными таблицами ББК на фоне появления разделов модернизированных средних таблиц обусловлена необходимостью сохранения и совершенствования средств обеспечения доступности бесценного фонда РГБ как культурного достояния страны в плане реализации тематического поиска. Индексы, приписанные миллионам книг, диссертаций, нотно-музыкальных и картографических изданий, сохраняются в том неизменном виде, в котором они применялись и применяются при систематизации. В интересах современного пользователя модернизирована лексика и некоторые элементы, касающиеся постсоветского пространства. Иерархические деревья строятся, естественно, из индексов ББК и их полных словесных формулировок.

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 (с учётом грамматики), а также непосредственно по индексам [2]. 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, выбирает тему, и система передаёт индексы ББК в ЭБ, где отыскиваются библиографические записи и полные тексты документов.

Учёт грамматики означает, что друг другу приравнены при поиске формы падежей и чисел существительных, прилагательных, причастий, а также некоторые формы словообразования.

Примеры:

Человек=человеком =люди=людей=человеческий=человеческие=человеческими.

Дети=детей=ребёнок=ребёнка=детсткий=детские, детство.

Замораживание=замораживанием=замораживаний=замороженном=заморожен=заморожена= замороженная=замороженную.

Таким образом, на запрос «морской фитопланктон» будет найден документ по теме:

Е082.351.401 Биологические науки -- Общая биология -- Общая экология и биогеография. Охрана живой природы -- Гидробиология -- Приспособление водных организмов к жизни в толще воды и на дне бассейнов -- Флора водоемов -- Фитопланктон -- Моря и океаны.

В сущности, система реализует функции виртуального систематического каталога, так как она автоматически «выстраивает» библиографические записи, содержащие индексы полных таблиц ББК за виртуальными разделителями, т.е. индексами и их словесными формулировками. Важно то, что при этом наличие или отсутствие «расшифровки» индекса в БЗ не имеет значения. Навигатор отбирает те записи, в которых индекс запроса совпадает с искомым или хотя бы с его начальными знаками.

Вернёмся к нашему примеру БЗ. При поиске по словам *«обмен веществ микроорганизмов»* или *«микроорганизмы, обмен веществ»* по базе данных навигатора в выдаче оказывается и наш автореферат, так как в его метаданных имеется соответствующий индекс ББК, хотя и без словесных формулировок.

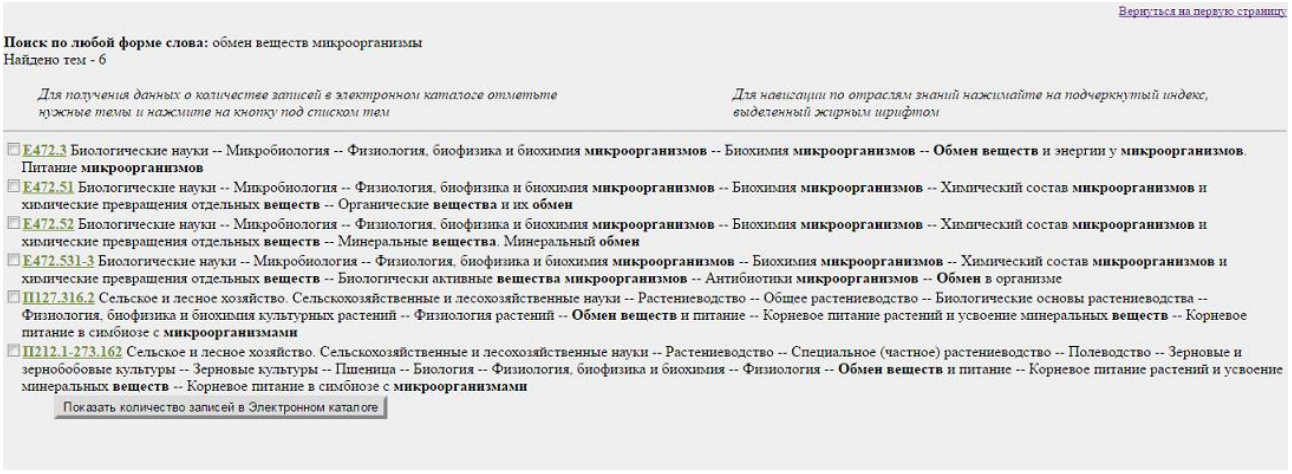

Рис.1. Экран результата поиска по словам

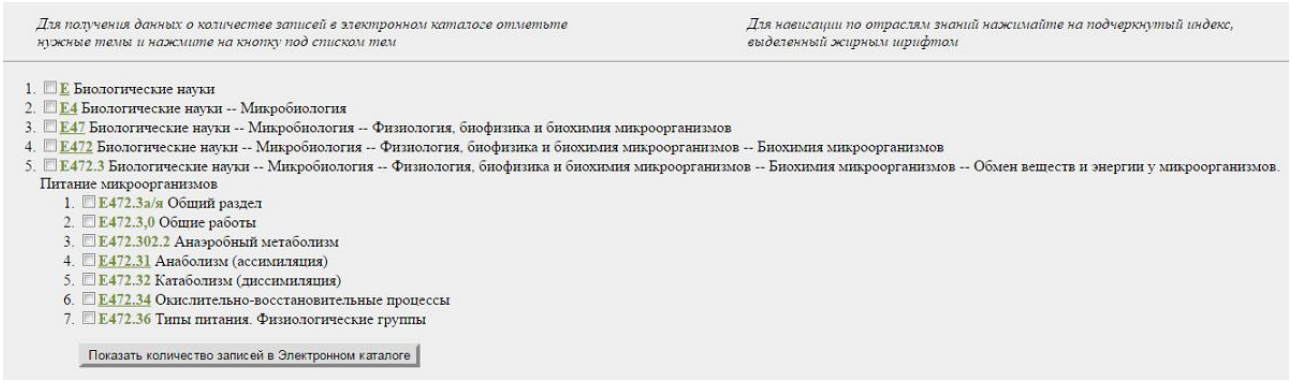

Рис. 2. Экран вывода фрагмента иерархического дерева для выбранной темы

Документ может быть найден и при выборе темы путём просмотра в навигаторе с верхнего уровня классификационной модели.

Обогащение запроса происходит за счёт присоединения к заданным словам их грамматических вариантов, а затем – дополнения найденного классификационного индекса нижестоящими индексами. От второго режима пользователь имеет возможность отказаться и получить только документы с верхним из найденных индексов.

3.2. Классификационная модель в среде LOD

Настала очередь для выполнения желаний пользователя относительно вывода ресурсов библиотек в открытое сетевое пространство: для опубликования нашей системы знаний в среде связанных открытых данных. Эта задача заключается, в первую очередь, в формировании отдельных утверждений, построенных в среде описания ресурсов RDF (Resource Description Framework) [3]. Ресурсами считаются любые данные, в том числе элементы классификации. Каждый ресурс получает URI (Uniform Resource Identifier, универсальный идентификатор ресурса в сети), т.е. уникальный адрес. Любое утверждение о ресурсе выглядит как триплет (тройка) «субъект – предикат – объект». Важно, что в технологии LOD требуется обеспечить процессы обогащения запроса поисковыми признаками исключительно с помощью программных средств (без участия человека). Это обстоятельство предъявляет особые требования к качеству структуры данных.

Проектные решения РГБ

Третий год специалисты РГБ ведут проект «Представление классификационных метаданных электронных библиотек по технологии связанных открытых данных (Linked Open Data)» [4, 5]. Он поддержан грантом РФФИ № 15-07-05265.

По сути, технологии и данные, реализованные в навигаторе (виртуальном систематическом каталоге) переводятся в пространство Semantic web. Файлы классификации, полученные в результате преобразования отредактированных разделителей систематического каталога РГБ в RDF, были успешно загружены в семантическое хранилище для последующего манипулирования данными с помощью языка запросов SPARQL. Программное обеспечение включает программный пакет, удовлетворяющий всем требованиям - Apache Jena, который является платформой для создания приложений связанных данных и Семантической паутины (подробнее [5]). В частности, помимо протокола SPARQL, сервер Fuseki (компонента платформы) поддерживает полнотекстовые запросы (Jena text query) к встроенному серверу Lucene.

На основе анализа зарубежного опыта преобразования в RDF других классификаций и иных семантических структур **выработана собственная концепция решения данной задачи.** Основное отличие в том, что наша модель знаний формируется не на основе некоего эталона классификации со всеми составляющими и правилами построения индексов персоналом библиотек, а на базе готовых индексов систематического каталога, уже построенных для конкретных документов. Естественно, и основные деления классификации из таблиц в иерархии присутствуют.

Каждый классификационный индекс объявляется концептом и получает URI. В форме триплетов представляются все связи индекса с теми элементами классификации, которые могут использоваться для программного обогащения запроса человека. Таковыми считаются: эквиваленты слов из формулировок (грамматические формы, результаты словообразования, синонимы и т.д.), иерархические и ассоциативные связи между индексами, ассоциативные связи с другими ресурсами в LOD.

Определены те пространства имён в сети, из которых берутся метки (тэги) [6, 7]: RDF, SKOS, RDFS.

Разработан 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 базы данных (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.).

Покажем процесс представления в RDF индекса Е472.311.5 из рассмотренного выше примера.

Вначале указываются те пространства имён в сети, из которых берутся метки (тэги) для описания данных (выделены полужирным шрифтом):

@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

@prefix rdfs: http://www.w3.org/2000/01/rdf-schema#

Затем для каждого индекса автоматически был создан концепт с URI, в том числе для индекса E472.311.5.

<*rdf:RDF xmlns:skos="http://www.w3.org/2004/02/skos/core# "*> – это метка начала утверждения (записи) о концепте (индексе классификации) в целом.

Далее указывается, что обозначением на естественном языке (notation) для данного URI является *E472.311.5*: <*skos:notation*> *E472.311.5*</*skos:notation*>.

Здесь <*skos:notation*> – это метка начала записи в RDF, а </*skos:notation*> – метка конца записи.

Далее используется аналогичная структура описания.

Затем фиксируется отношение[1] *«E472.311.5* имеет полную цепочку формулировок *Биологические науки -- Микробиология -- Физиология, биофизика и биохимия микроорганизмов -- Биохимия микроорганизмов -- Обмен веществ и энергии у микроорганизмов. Питание микроорганизмов -- Анаболизм (ассимиляция) -- Биосинтез -- Фотосинтез. Фотосинтезирующие микроорганизмы»:*

<skos:prefLabel xml:lang="ru"> *Биологические науки -- Микробиология -- Физиология, биофизика и биохимия микроорганизмов -- Биохимия микроорганизмов -- Обмен веществ и энергии у микроорганизмов. Питание микроорганизмов -- Анаболизм (ассимиляция) -- Биосинтез* </skos:prefLabel>.

Далее фиксируем другие отношения данного индекса, в частности, отношение:

«*E472.311.5* имеет вышестоящий индекс *E472.311»:*

<skos: broader xml:lang="ru"> *E472.311*</skos: broader>.

<skos: </rdf:RDF> – метка окончания утверждения относительно данного концепта в целом.

В рамках проекта создаются дополнительные средства обогащения запросов пользователей, например, программно вносятся в RDF-представления индексов поисковые слова из методических указаний к ним, взятых из таблиц ББК. Они связываются с концептом с помощью метки skos:example.

Например, для индекса *Щ314.043 (Искусство. Искусствознание -- Музыка -- Отдельные виды музыки и музыкального исполнения -- Вокальная музыка -- Теория вокальной музыки -- Виды, жанры и формы вокальной музыки -- Вокальные жанры камерного репертуара*) кодируется методическое указание: «*Кантата, вокальный цикл, романс, песня, баллада и т. п.».*

Таким образом, формируются данные в среде LOD. Это позволяет связать в Семантической паутине что угодно с чем угодно, а также обеспечить поиск связанных открытых данных стандартными программными средствами сети с обогащением запросов на основе зафиксированных связей.

4. Связанные данные в Semantic web

Структурированные данные нетрудно преобразовать программно в RDF-представления. Более сложным оказывается выбор ресурсов, с которыми имеет смысл соединять классификационную модель знаний. Эту задачу решают многие библиотеки в мире.

Рассмотрим применение классификационной модели в технологии LOD в двух направлениях. С одной стороны, ясно, что публикация классификационной модели в пространстве связанных открытых данных позволит обогатить запросы различных пользователей (организаций и лиц) для программной передачи этих запросов в ЭК и ЭБ РГБ. Если другие библиотеки найдут эту технологию полезной для себя, то также смогут её использовать для обогащения запросов в их ЭК. Более того, построение (в форме RDF-высказываний) связей индексов полных таблиц из разработанной модели и индексов средних таблиц, которые будут представлены в такой же форме, позволило бы

[1] В реальных RDF- записях вместо индексов *E472.311.5и E472.311* указываются их URI.

обогащать запросы пользователей ещё и индексами последних для передачи запросов в ЭК, работающих на среднем варианте таблиц ББК.

Среди различных вариантов использования классификационной модели рассматривается двусторонняя связь с Википедией, которая очень популярна практически у всех разработчиков LOD для библиотек. С одной стороны, можно предлагать пользователям нашей ЭБ предоставление статей из Википедии для заданных ими в запросе терминов или названий. С другой стороны, служба поддержки Википедии могла бы, если заинтересуется, посылать термины в нашу систему на поиск публикаций по соответствующей теме в ЭК РГБ и сообщать своим пользователям на сайте о наличии публикаций по данной теме в Библиотеке. Собственно, такого рода возможности – предмет дальнейших исследований.

Публикации

1. Лавренова О.А. Семантические средства библиографического поиска в Российской государственной библиотеке. // Общетеоретические и футурологические проблемы библиографии. Библиографическая запись как основа формирования библиографических ресурсов : материалы II Международного библиографического конгресса «Библиография: взгляд в будущее (Москва, 6-8 октября 2015 г.) / Рос. гос. б-ка. Москва: Пашков дом, 2016. С. 309–323.
2. Лаврёнова О.А. Представление классификации в Семантической паутине. // Информационное обслуживание в век электронных коммуникаций: XI Всероссийская научно-практическая конференция «Электронные ресурсы библиотек, музеев, архивов», 2–3 ноября, 2016 г., Санкт-Петербург: сборник материалов. – СПб: 2016. С. 73–92.
3. RDF 1.1 Concepts and Abstract Syntax. URL: https://www.w3.org/TR/rdf11-concepts/
4. Шварцман М.Е., Найдин О.П. Linked Open Data как средство обогащения поисковых запросов // Университетская книга .– 2015 .– №12. С. 66–71.
5. Лаврёнова О.А., Павлов В.В. Библиотечно-библиографическая классификация как традиционная система организации знаний в среде открытых связанных данных / Научно-технические библиотеки. – 2017. – №4. – С. 44–46.
6. Namespaces in XML 1.0 (Third Edition). URL: https://www.w3.org/TR/xml-names/
7. SKOS. Simple Knowledge Organization System Primer. W3C Working Group Note 18 August 2009. URL: http://www.w3.org/TR/skos-primer/