Лавренова Ольга Александровна

Электронные библиотеки и тематический поиск в среде Semantic WEB на основе связанных LOD – словарей

1. РГБ опубликовала Классификационную систему организации знаний, созданную на основе отечественной классификации ББК и Генерального систематического каталога (ГСК). Система в настоящее время функционально связана с электронным каталогом (ЭК) и электронной библиотекой (ЭБ) РГБ. Путём обогащения запросов пользователей на основе семантических связей между темами (рубриками) Классификационная система поддерживает тематический поиск, реализуя функции виртуального систематического каталога. Она опубликована в среде связанных открытых данных (Linked Open Data, LOD) под открытой лицензией РГБ, основанной на рекомендациях Creative Commons, опция CC BY-NC-ND «Attribution-NonCommercial-NoDerivs» (полномочия – некоммерческое использование – без права изменения). Проект Классификационной системы был разработан при поддержке РФФИ (проект 15-07-05265).

2. Настоящая статья подготовлена по материалам доклада на Двенадцатой ежегодной научно-практической конференции Российской ассоциации электронных библиотек (октябрь 2018 г.). Запись трансляции конференции размещена на сайте Ассоциации.

3. Проверив работу Классификационной системы на реальном поиске электронных ресурсов РГБ, мы поставили задачу **предоставления в пользование различным библиотекам созданных возможностей тематического поиска**, включая использование при обогащении исходных запросов для дальнейшего поиска в других библиотеках:

– иерархических и ассоциативных связей между темами,

– форм словоизменения и словообразования используемых слов русского языка;

– отношения синонимии (в дальнейшем).

Считаем целесообразным реализовать тематический иерархический поиск на основе нашей Классификационной системы и для Национальной электронной библиотеки (НЭБ), но это отдельный вопрос.

Сетевые системы организации знаний (Networked Knowledge Organization Systems, NOS), средства обогащения тематических запросов пользователей и концепция построения нашей Классификационной системы уже рассматривались ранее в 2-х публикациях «Публикация в среде Linked Open Data классификационной системы организации знаний» и «Публикация в среде Linked Open Data классификационной системы организации знаний (часть 2)».

4. В чём, собственно, проблема? Для проведения тематического поиска библиотеки в нашей стране используют в своих автоматизированных информационных системах (АИС) разнообразные информационно-поисковые языки (ИПЯ): библиотечно-библиографические классификации (преимущественно различные варианты ББК, а также УДК), системы предметных рубрик, тезаурусы, свободные ключевые слова. При этом отсутствуют средства установления соответствий между их элементами, не говоря о совместимости различных ИПЯ. Пользователи электронных каталогов или электронных библиотек поставлены перед фактом: проводить поиски с помощью средств конкретной АИС и довольствоваться тем качеством (полнотой и точностью получаемой выдачи), которое эта система способна обеспечить. Большинство пользователей над данным обстоятельством не

задумываются, хотя оно приводит к потерям информации, причём, неизвестным человеку, или, напротив, к выдаче ему слишком большого процента нерелевантных публикаций.

В то же время, АИС ряда других библиотек способны обеспечить на том же массиве документов поиск с существенно более высоким качеством. Такая ситуация сохраняется **более 50-ти лет** развития автоматизированного информационного поиска.

Полнотекстовый поиск в целых документах, аннотациях или рефератах в ряде систем даёт хорошие результаты, но **не позволяет гарантировать** учёт при поиске понятий, более узких, чем заданное в запросе пользователя, или связанных с ним ассоциативными отношениями.

Энциклопедический поиск, используемый, например, в НЭБ, на данный момент недостаточно изучен и не отлажен, хотя представляется перспективным.

Проводимые в настоящее время разработки, которым был посвящён доклад, направлены на обеспечение **информационного равенства граждан в плане получения качественных результатов поиска по темам** (предметам). Планируется создать для пользователей возможности обогащения их запросов дополнительными поисковыми признаками, зафиксированными в Классификационной системе организации знаний с программным переходом к поиску по запросам, обработанным таким образом, в АИС других библиотек.

5. Предлагается следующий **путь движения** к относительному информационному равенству пользователей электронных библиотек и каталогов в плане тематического поиска:

– представление отдельных лексических элементов классификаций, файлов предметных рубрик, тезаурусов в среде связанных открытых данных и формирование так называемых «словарей связанных открытых данных» (Linked Open Data Vocabularies, LOD Vocabularies, LOD-словарей);

– разработка способов формирования связей между лексическими элементами различных словарей для тематического поиска на основе исследования их лексического материала;

– формирование единой точки доступа к поиску по различным словарям на базе Классификационной системы организации знаний; предоставление, тем самым, в пользование различным библиотекам возможностей тематического поиска с учётом иерархических и ассоциативных связей между темами, а также словоизменения и словообразования используемых слов русского языка (семантическое обогащение исходных запросов);

– реализация программного выявления соответствующих по смыслу поисковых элементов в других LOD-словарях для поиска по запросу, которые был отправлен в Классификационную систему; переход с использованием таких LOD-словарей в определённые АИС, которые используют соответствующие словари для описания смыслового содержания публикаций.

Можно сказать, что в рамках предлагаемого проекта планируется разработать систему, формирующую семантическое пространство организации знаний для поддержки тематического поиска в электронных ресурсах библиотек, использующих различные ИПЯ.

6. **Основной целью доклада было формирование сотрудничества** с другими библиотеками по следующим задачам построения семантического пространства связанных открытых данных: – публикация в Semantic Web Среднего варианта таблиц ББК, а также нормативного файла (тезауруса) географических названий; – создание связей Классификационной системы на основе ББК/ГСК с этими ресурсами в среде LOD; – формирование связей ресурсов Системы с опубликованными в среде LOD ресурсами УДК;

– изучение возможностей формирования связей Системы с фрагментами национального файла предметных рубрик РНБ и тезаурусом по медицине MeSH (Medical Subject Headings).

В результате, нам удалось обнаружить коллег (сначала на Конгрессе РБА, а теперь и на Конференции ЭЛБИ), которые выразили желание участвовать в нашем проекте. Это большая удача.

7. На рисунке 1 представлена упрощённая схема проектируемого процесса использования **связей между лексическими единицами различных LOD-словарей** для обогащения запросов при использовании разных ИПЯ.

Допустим, запрос поступает в нашу Классификационную систему организации знаний. Она обогащает его всеми грамматическими формами составляющих его слов (например, *ребёнок-дети-детей, психология-психологию*), некоторыми формами словообразования (например, *дети-детский-детство*), словами из словесных формулировок индексов делений более низких уровней иерархии или ассоциативных делений (связь «смотри также»). Целесообразно в будущем решить проблемы учёта синонимии и условной эквивалентности лексических единиц (например, присоединять к слову «*языкознание*» в запросе слов «*лингвистика*» и «*языковедение*».

Рис. 1. Упрощённая схема использования семантических связей между ресурсами, представляемыми в среде LOD

Затем на основе такой обработки Система найдёт по словам словесные формулировки и, соответственно, индексы полных таблиц ББК, построенных при каталогизации. Эти индексы передаются в ЭК РГБ для поиска библиографических записей. Если имеются географические названия, то предлагается с помощью обращения в соответствующий тезаурус добавить связанные формы названий тех же географических объектов, как это делается и в обычном ЭК.

Мы предлагаем передавать обогащённый запрос в другие электронные каталоги, например, работающие на основе Средних таблиц ББК, УДК или тезауруса MeSH. Для этого требуется обогатить запрос теми индексами или дескрипторами из указанных систем организации, которые соответствуют по смыслу выбранным индексам полных таблиц ББК из нашей Классификационной системы, а затем отправить обогащённый запрос в чужой каталог.

8. Итак, планируется сформировать новые LOD-словари. Все лексические единицы классификаций, файлов предметных рубрик, тезаурусов потребуется опубликовать в среде LOD как концепты. Фактически обрабатываемые данные из каждого словаря структурируются в RDF (Resource Description Framework) на основе модели (пространства имён) SKOS (Simple Knowledge Organization System, Простая система организации знаний), предназначенной для создания именно такого рода словарей связанных данных. Каждый элемент данных получит URI (Uniform Resource Identifier, универсальный идентификатор ресурса в сети), т.е. уникальный адрес. Между URI лексических единиц словаря, а затем и единицами различных LOD-словарей устанавливаются смысловые связи. Примеры представления ББК в RDF имеются, в частности, в наших публикациях http://www.gpntb.ru/win/inter-events/crimea2017/disk/019.pdf и www.unkniga.ru/innovation/tehnology/7808-tematicheskiy-poisk-elektronnyh-resursov-na-osnove-klassifikatsionnoy-modeli.html,

9. Ключевой и наиболее сложной является **проблема установления степени семантического соответствия (смыслового сходства) между делениями, рубриками, дескрипторами, ключевыми словами как коротенькими текстами в различных поисковых LOD-словарях**. Полагаем, что основная сложность заключается именно в структурных особенностях информационно-поисковых языков, лексические единицы которых формируются по различным правилам. В частности, тезаурусы фиксируют только парадигматические связи (отношения), а классификации УДК, ББК и предметные рубрики – одновременно парадигматические и синтагматические.

Пример индекса с цепочкой словесных формулировок:

Ш141.2-032я721 *Филологические науки. Художественная литература — Языкознание* -- *Индоевропейские языки — Славянские языки — Восточнославянские языки — Русский язык* -- *История языка — Историческая лексикология — Этимология*

Словесные формулировки, выделенные зелёным цветом (иерархия филологических наук), находятся в парадигматических отношениях, как это всегда бывает в тезаурусах. Словесные формулировки, выделенные синим цветом (иерархия индоевропейских языков), находятся тоже в парадигматических отношениях. Однако между формулировками зелёного и синего цвета существуют синтагматические отношения.

Следует отметить, что в большинстве случаев **прямые смысловые связи** между ресурсами различных LOD-словарей построить будет невозможно. Поэтому предварительно в ходе экспериментов по установлению соответствий индексов УДК и ББК мы использовали три

типа связей между конкретными элементами: **«точное соответствие», «близкое по значению», «имеет отношение к теме».**

Исследования и многочисленные эксперименты должны проводиться на достаточно представительном материале из различных работающих классификаций, файлов предметных рубрик и тезаурусов, преобразованных в LOD-словари. В частности, необходимо получить и преобразовать файлы индексов Среднего варианта таблиц ББК, построенных при обработке поступающих в фонды библиотек документов. Об этом тоже удалось договориться с коллегами из других организаций.

10. **На конференции также обсуждались вопросы** о состоянии этого направления за рубежом, о необходимом финансировании проекта и т.д.

Была также **продемонстрирована технология работы Классификационной системы на основе ББК\ГСК**, но лучший способ ознакомления с системой – это реальная работа по поиску в её базе данных. На рисунке 2 изображён экран входа в Систему. В разделе «О проекте» приведена подробная инструкция по работе с системой с примерами, а также полное описание проекта. Поиск можно начать, в частности, путём ввода произвольного сочетания слов запроса в поисковое окно. Другой вариант поиска – навигация в разделе «Классификационная система» по иерархическому дереву делений и с использованием переходов по связям «смотри также». В дерево можно перейти и из режима поиска по словам. Результат – сначала список тем (делений), затем – перечень библиографических записей в ЭК РГБ со связями с ЭБ.

Раздел «Экспериментальная система» демонстрирует экспериментальные примеры частично автоматизированного установления соответствий индексов ББК и УДК (около 1000 разделов УДК), а также переводов на английский язык словесных формулировок индексов раздела Щ «Искусство. Искусствознание» (более 6000).

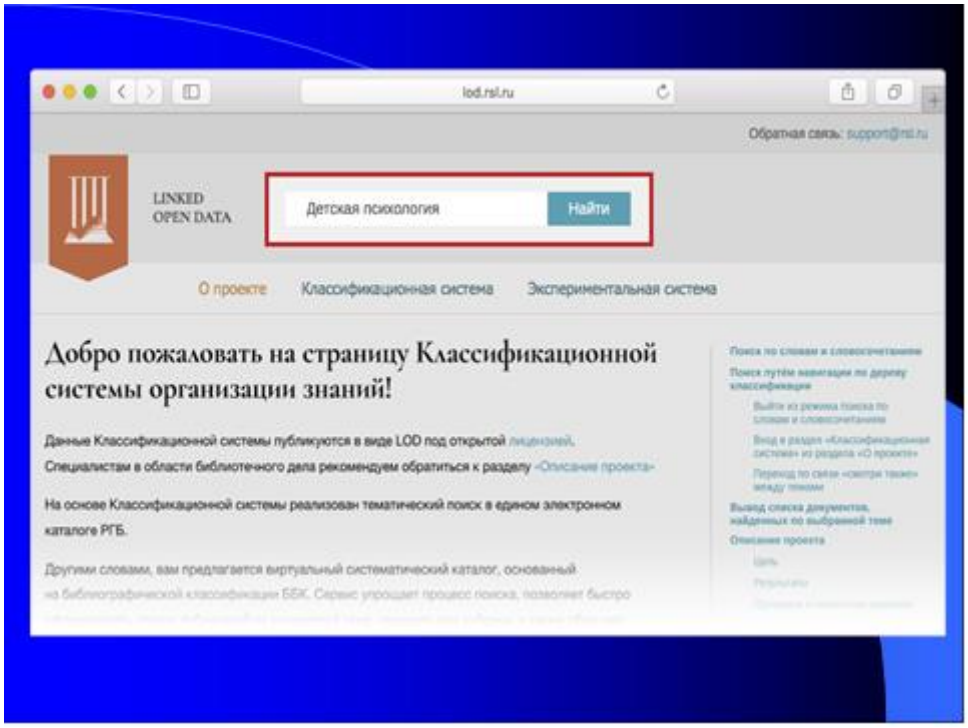

Рис.2. Первая страница Классификационной системы