Программа представления библиографических данных в Семантическом Webe.

I

Призрак Семантического Weba

Сегодняшний наш доклад – это продолжение серии докладов, посвящённых новому библиографическому окружению, которые озвучивались на конференции ЛИБНЕТ в течение ряда последних лет. Мы будем говорить сегодня исключительно о перспективах участия библиотек в Семантическом Webe, как о теме давно назревшей, предполагая, однако, иной эмоциональный накал, поскольку, что бы мы ни говорили, а робкий вклад российских библиотекарей в наше общее будущее – имеется в виду скромность только в означенном контексте, безусловно, - не идёт ни в какое сравнение ни с масштабами задачи, ни с участием зарубежных библиотек в процессе обустройства этого будущего.

Мы предлагаем всё-таки перейти от слов к делу, и в связи с этим решили изложить наш взгляд на то, что должно делаться и как это должно делаться для развития качественно новой организации библиотечных данных и внедрения их в Семантический Web.

Коротко о терминологии. Есть несколько базовых терминов, связываемых с Семантическим Webом. Собственно Семантический Web (Semantic Web) (или Семантическая паутина), или Web данных, Связанные данные (Linked Data), RDF. Иногда эти термины используют как синонимы, и это вносит определённую путаницу. Скорее, в перечисленном порядке они означают движение от общего к частному.

Семантический Web – это интернет-среда, в которой данные размещаются таким образом, что их смысл становится доступным для машинной обработки.

Связанные данные – это основная идея организации Семантического Weba, но всё же лишь один из возможных способов его реализации.

В своё время мы предлагали, и в части формата RUSMARC даже осуществили, идею размещения в Интернете стандартов так, чтобы данные этих стандартов могли обрабатываться непосредственно компьютерами. Это в полной мере соответствует определению Семантического Weba, но никакого отношения не имеет к Связанным данным.

RDF– конкретная модель, используемая для связи данных.

Часто доклады, посвящённые семантическому Weby и Связанным данным украшают диаграммой из Linked Open Data dataset catalog, на которой связанные данные образуют облако (Linked Open Data Cloud) связанных пузырей, где пузыри – это наборы данных. Мне кажется, что больше смысла в аналогии Связанных данных с мозгом, не уверен, что даже с человеческим, где понятия, хранимые в отдельных ячейках памяти, ассоциативно связываются друг с другом. Разница вероятно главным образом в том, что мозг вряд ли хранит образы в том виде, в котором их воспринимаем мы, и наше восприятие образуется

в основном за счёт ассоциативных связей. Однако и Семантический Web, возможно когда-нибудь, придёт к тому же.

Это не просто лирическое отступление, это к вопросу о философской значимости начинания консорциума W3C. Продолжая нашу аналогию: мозг ведь не хранит полные тексты, а образует их (если вообще образует) благодаря ассоциативным связям.

17 мая 2001 года в журнале «Scientific American» вышла статья Тима Бернерс-Ли, Джеймса Хендлера и Оры Лассилы «A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities» (Новая форма содержания Сети, понятная компьютерам, произведет революцию в ее возможностях). Эту дату многие считают датой рождения Семантического Weba. Хотя идея вынашивалась консорциумом W3 по-видимому ещё с середины 90-х годов прошлого века, откуда появились документы, такие как RDF Model and Syntax (1999 Recommendation) и RDF Schema (2000 CandidateRecommendation), посвященные использованию модели Resource Description Framework (RDF) для обработки метаданных. 10 февраля на сайте консорциума W3C опубликован так называемый «набор из шести» (set of six) документов: (Primer, Concepts, Syntax, Semantics, Vocabulary, и TestCases). Это основа Семантического Weba. Тогда же, 10 февраля 2004 г. на сайте W3C появляется руководство по "OWL" (Web Ontology Language) – языку описания онтологий. С этого момента развитие Семантического Weba идёт уже полным ходом. Разрабатывается необходимая документация и инструментарий: соответствующие поисковики и браузеры, редакторы, средства тестирования и публикации данных. Разрабатываются отраслевые онтологии для торговли, различных отраслей промышленности, для издающих организаций, для различных отраслей науки, для социальных сетей.

Зарубежные библиотеки также не остаются в стороне от прогресса. В настоящее время в виде связанных данных уже опубликован ряд наборов библиографических записей (национальных библиографий и каталогов): Британская национальная библиография, Французская национальная библиография, Немецкая национальная библиография, OCLC WorldCat, Шведская национальная библиография. Отдельный разговор будет о том, какую работу в этом направлении ведёт Библиотека Конгресса. Это наверняка неполный перечень библиографий, представленных в виде Связанных данных, здесь перечислены только те участники процесса, информация о которых была доступна автору на момент подготовки доклада.

В прошлогоднем докладе О.Н. Жлобинской "Semantic web, связанные данные и библиотеки" есть перечень традиционных библиотечных словарей, опубликованных в виде Связанных данных. Процитирую:

«• Словари RDA (OMR, 2011) –плоские словари (перечни терминов) без иерархических связей.
• LCSH (Library of Congress Subject Headings) (http://id.loc.gov/)
• FAST (Faceted Application of Subject Terminology)
• MESH (Medical Subject Headings)
• Form and genre headings for fiction and drama
• TGM (Thesaurus for Graphic Materials)

• SWD (Немецкиепредметныетермины)
• RAMEAU (Французскиепредметныерубрики)
• UDC summary (около 2400 классов УДК на 40 языках (http://udcdata.info)
• VIAF (Virtual International Authority File) – опубликованный OCLC виртуальный международный авторитетный файл - множество связанных контролируемых словарей (авторитетные записи на имена лиц) национальных библиографирующих агентств».

Публикация традиционных словарей в виде Связанных данных – тема, заслуживающая отдельного обсуждения и отдельной разработки, благодаря тому мощному потенциалу, который содержит это направление деятельности.

II

Что это даёт?

Мы кратко обрисовали то, что происходит в мировой информационной среде в отношении Семантического Weba. Давайте попробуем разобраться, а зачем это в ней происходит?

В упомянутом докладе О.Н. Жлобинской выделены 7 существенных достоинств использования Связанных данных. Я бы рекомендовал обратить внимание на этот список, поскольку, на мой взгляд, он абсолютно справедлив. Со своей стороны хочу выделить три основных момента, которые определяют необходимость Семантического Weba для библиотек.

Первое, и это совпадает с одним из пунктов списка О.Н. Жлобинской: библиотечные метаданные становятся доступными для поисковых машин в Webe – ссылки к библиотечным коллекциям могут устанавливаться извне и в форме, понятной для поисковых машин.
В одном из своих докладов несколько лет назад я ссылался на данные исследования OCLC, 2004-го года, по-моему, о том, что только 20% пользователей, желающих найти в Интернете ту или иную книгу, обращаются к каталогам библиотек. Т.е. 80% потенциальных читателей удовлетворяются результатами, даваемыми поисковыми машинами, и для библиотек потеряны, так же, как и библиотеки для них. Думаю, за последние годы, за счёт увеличения Интернет-контента, это соотношение существенно изменилось, и не в пользу каталогов библиотек.
Делая доступными библиотечные метаданные для поисковых машин, библиотеки таким образом просто присоединяют свой ресурс к общему Web-ресурсу, по которому те осуществляют поиск.

Это очень важное обстоятельство, но вряд ли решающее, поскольку в принципе возможно сделать метаданные доступными для поисковых машин и без Семантического Weba.

Второе: новые возможности создания и объединения авторитетных данных: авторитетные записи могут не храниться в едином файле, а собираться «на лету» из разных авторитетных источников. Это означает, что и создание того или иного международного авторитетного файла может осуществляться простым указанием ссылок на соответствующие заголовки.

Третье: процитирую упомянутую выше статью Тима Бернерс-Ли и др. «Новая форма содержания Сети, понятная компьютерам, произведет революцию в ее возможностях», ставшей началом Семантического Weba:

«Такие паутино-подобные системы [речь идёт о системах семантически связанных данных] предлагают массу удивительных возможностей всем, начиная от крупных компаний и заканчивая обычными пользователями, и дают такие преимущества, предсказать которые заранее трудно или даже невозможно».

Это определение может относиться вообще к любому прогрессивному начинанию, и в этом-то и кроется самое главное. Главное то, что, присоединяясь к общему движению в сторону развития Семантического Weba, библиотеки встают на путь прогресса, тот же, каким развивается всё мировое информационное сообщество, и получают в перспективе доступ к «массе удивительных возможностей» и преимуществ, «предсказать которые заранее трудно или даже невозможно».

И закончу этот краткий анализ ещё одной цитатой из той же статьи:

«В полную силу Семантическая Сеть будет реализована тогда, когда люди создадут множество программ, которые, знакомясь с содержимым Сети из различных источников, обрабатывают полученную информацию и обмениваются результатами с другими программами. Эффективность таких программных агентов будет расти экспоненциально по мере увеличения количества доступного машинно-воспринимаемого веб-контента и автоматизированных сервисов (включая других агентов)». Пер. Е. Золина.

III

А что же в России?

Должен оговориться, что речь пойдёт только о библиотеках, поскольку есть существенная разница между реакцией в обсуждаемом аспекте информационного и научного сообществ России – без библиотек – и реакцией самих библиотек.

Пока приходится констатировать, что мощные ветры перемен, весьма ощутимые в мировом библиотечном сообществе, в российской его части вызывают лишь лёгкую рябь на поверхности.

Я в частности вот о чём. Известно, что IFLA, например, формирует пространства имён RDF для областей ISBD. Российские библиотекари также принимают участие в этой работе. Это даёт возможность думать, что наше библиотечное сообщество не стоит в стороне от новых веяний, а движется в общем потоке к Семантическому Weby, Связанным данным, RDFи прочему из того же ряда. Вопрос, а кто-нибудь из принимающих участие в этой работе понимает, зачем это делается? Не хочу обидеть наших библиотекарей, но берусь утверждать, что и сами вдохновители идеи создания пространств имён для ISBDдо конца не понимают, что из этого должно получиться.

Интересно, что представители тех же библиотек, которые в IFLA создают пространства имён для ISBD, одновременно создают пространства имён для своих национальных библиографий. Казалось бы, зачем? Не лучше ли, если мы работаем по правилам ISBD,

завершить один коллективный труд в отношении этого стандарта, и всем потом пользоваться его результатами. Не ждёт ли российское библиотечное сообщество чего-то подобного? Но ведь аналогичная загадочная история происходит с пространствами имён RDA и пространствами имён Библиотеки Конгресса и Британской библиотеки. Это совершенно разные пространства имён.

И на самом деле понятно, почему это так. Дело в том, что модель RDF оперирует таким понятием как ссылка на ресурс. Через такие ссылки определяются объекты, свойства и, как правило, субъекты выражений RDF. Совокупности таких ссылок и образуют пространства имён для выражений RDF. Если наша мудрая позиция «ты, барин, запевай, а мы подхватим» принесёт желаемые плоды, и мы сможем дождаться лучшего из созданного другими, то существование нашей национальной библиографии в RDF будет сильно зависеть от возможности доступа к выбранным пространствам имён. Вы знаете, что в октябре этого года был закрыт доступ к серверу Библиотеки Конгресса loc.gov. Если бы наша национальная библиография в RDF основывалась на пространствах имён Библиотеки Конгресса, то на это время, или на любое другое, по выбору Библиотеки Конгресса, она (наша национальная библиография) превратилась бы в набор бессмысленных инструкций.

Ещё один аргумент в пользу создания национальных наборов пространств имён. В случае, если российское библиотечное сообщество будет разрабатывать собственные наборы пространств имён для представления библиографических данных в RDF, это даст определённый приоритет для развития отечественных библиотечных программных продуктов, предназначенных для работы с такого рода данными.

При этом, при корректном определении терминов национальных словарей (пространств имён) с указанием необходимых ссылок на стандартные классы или свойства, на термины словарей ISBD, RDF, FRBR и на термины других национальных словарей, использование собственных словарей не создаст таких проблем, какие в настоящее время создаёт работа в разных форматах. Скорее всего, вообще не создаст никаких проблем.

IV

Что же дальше?

После того как прошёл первый порыв воодушевления, и наиболее продвинутые библиотеки мира отчитались о представлении национальных библиографий и других библиотечных ресурсов в Семантическом Webe, оказалось, что содержание необходимых для этого словарей сильно зависит от принятой модели исходных данных. Стало очевидным, по крайней мере для крупнейших национальных библиотек, что корректная разработка пространств имён, необходимых для представления национальных библиографий в Семантическом Webe, должна исходить из чёткого представления, что же должно получиться в результате.

А именно:

Что мы собираемся публиковать – полные библиографические записи, или некий урезанный массив, связанный с полными библиографическими данными – как своеобразный индекс к библиотечному каталогу?

Публикуем ли мы в виде связанных данных только библиографические данные, или публикуем и связываем с библиографическими данными также авторитетные и холдинговые данные?

Следующий вопрос: В Семантическом Webe забываем ли мы о накопленном опыте представления данных, выраженном в моделях, таких как FRBR? Если нет, то модель связанных данных, которую мы собираемся представить, должна удовлетворять и этим функциональным требованиям к библиографическим записям.

Это только примерный список вопросов, на который призваны ответить модели, разрабатываемые такими организациями, как OCLC, Библиотека Конгресса, Британская библиотека.

Вообще, мониторинг того, что делается мировым библиотечным сообществом в отношении Семантического Weba и понимание – на определённом уровне – цели, которой необходимо достигнуть даёт следующую картину того, что должно быть сделано нами, в России, и на каком уровне.

V

Что делать?

Прошу относиться к списку задач и особенно к распределению обязанностей, приводимым ниже как к выражению личного, но обоснованного, мнения автора доклада.

1. Необходимо разработать, рассмотреть и принять базовую модель взаимосвязи библиографических сущностей.
2. На основе этой модели необходимо создать пространства имён для классов и свойств, используемых при описании библио-сущностей.

Уровень решения этих двух задач. Если исходить из существующего положения вещей, то обе эти задачи относятся к компетенции Межведомственной комиссии по Каталогизации, поскольку их решение – есть по сути **создание правил каталогизации в Семантическом webe.** Правил, которые с большой долей вероятности со временем заменят существующие правила. В любом случае – это задачи, которые должны решаться на национальном уровне внутри библиотечного сообщества. Последнее – внутри библиотечного сообщества – тоже не проходная фраза, поскольку не исключён и другой вариант. Продолжение тактики пассивного выжидания может привести к тому, что эти вопросы будут решены информационным сообществом, - без особого учёта накопленного библиотекарями опыта и библиотечных традиций. Информационное сообщество уже и в России весьма активно занимается вопросами семантического weba.

Однако, надо отдавать себе отчёт, что решение обозначенных выше вопросов потребует от членов комиссии по каталогизации качественно новых знаний и навыков и радикального изменения видения предмета их деятельности, в силу осуществления её в

качественно новой среде. Возможно, с появлением специализированных интерфейсов это будет и не совсем так, но сейчас составление библиографических записей, с технической точки зрения, ближе к программированию, чем к привычному образу действий каталогизаторов или библиографов. В то же время эта операция, как и прежде, потребует использования накопленного веками опыта описаний объектов в качественно новом аспекте.

3. Третья задача – это задача размещения пространств имён классов и свойств, используемых библиотечным сообществом. В то время как отработка взаимодействия и проверка работоспособности создаваемых многосвязных систем может проводиться на серверах любых заинтересованных организаций, окончательный вариант пространств имён в силу их особой важности должен быть размещён на мощном сервере, отвечающем требованиям надёжности и обеспечивающим необходимый трафик. Вообще под задачу размещения библиографических данных в Семантическом Webe должна быть создана соответствующая информационная инфраструктура, обеспечивающая техническую сторону организации хранения и доступа к национальному библиографическому ресурсу.

   Решение этой задачи, естественно, должно обеспечиваться техническими специалистами, обладающими необходимыми знаниями в рассматриваемой области.
   Уровень решения: задача безусловно должна решаться на национальном уровне, либо в рамках деятельности МК РФ, либо с привлечением ресурсов Министерства связи и массовых коммуникаций, но, если мы ценим библиотечный опыт, постановка задачи и координация работ должны обеспечиваться Министерством культуры.
4. Четвёртая задача – преобразование накопленных библиотечных данных в новое качество.
   Уровень решения: составление таблиц соответствия полей формата RUSMARC терминам принятых пространств имён, – задача по природе своей относящаяся к сфере деятельности Национальной Службы развития системы форматов RUSMARC.
   Что касается самого конвертирования, оно должно проводиться самими владельцами информации в рамках соответствующих проектов МК РФ.
5. Не последняя по значению задача – подготовка специалистов, которая потребует основательного пересмотра содержания курсов обучения учебных заведений.
   Уровень решения: задача должна решаться как в рамках регулярного обучения на библиотечных факультетах и специализированных учебных заведениях, подведомственных МК РФ, так и на специализированных семинарах и курсах повышения квалификации.
6. Координация работ в целом. Повторяю, я высказываю своё личное мнение: поскольку все перечисленные задачи должны решаться на национальном уровне, а постановка задачи в целом – это перевод и дальнейшее поддержание массивов информации, накопленных и накопляемых библиотеками страны, в информационную среду web 3.0, то задача должна координироваться

Министерством культуры в целом, с ответственным исполнением, возложенным на Отдел библиотек МК РФ.

В любом случае, на это очень хотелось бы обратить внимание Отдела Библиотек МК РФ, который всегда являлся вдохновителем и организатором государственной библиотечной политики, и поддержку которого мы постоянно ощущали. Задача настоящего выступления и последующих – для каталогизаторов, - показать необходимость и неотвратимость предстоящих перемен, показать, что есть ресурс научный и интеллектуальный, который безусловно не исчерпывается докладчиками нашей конференции, - пока есть, - есть ощутимый задел, который может служить вполне конкретной основой для начала работы, есть, в конце концов, желание гордиться своей деятельностью. Чего же не хватает для того, чтобы поднять работу российского библиотечного сообщества в плане информационных технологий на мировой уровень? Что ещё нужно от нас, рядовых исполнителей?

На этой патетической ноте, я хотел бы завершить своё сегодняшнее выступление, - и это было бы правильно, - но должен сказать несколько слов о текущей деятельности Национальной Службы развития системы форматов RUSMARC.

VI

О MARC форматах вообще и о RUSMARCе в частности.

В конце октября этого года на форуме проекта BIBFRAME – вы знаете, это проект Библиотеки Конгресса по переходу к RDF представлению библиографической информации – появилось следующее эссе: «MARC is dead! Как много раз вы слышали это за свою библиотечную карьеру? На этот раз, однако, это действительно так!» И далее следовал анонс самого проекта BIBFRAME и двух Free BIBFRAME вебинаров. Не очень понятно, зачем это могло бы потребоваться участникам форума BIBFRAME, важно другое. На это немедленно последовал ответ представителя Библиотеки Конгресса Роя Теннанта: «Надеюсь, Вы не часто слышали это за свою карьеру, поскольку это искажение цитаты и неправда. В 2002 году я сказал "MARC Must Die" -- NOT "MARC is dead", и с тех пор эта цитата часто искажалась, MARC безусловно не умер, хотя я и могу доказать, что то, что я сказал – правда. Но это совершенно разные вещи». Привожу эту переписку потому, что и у нас в России уже приходится слышать подобное. Вот, например, почти цитата о впечатлениях одного из современных руководителей о Всемирном Библиотечном Конгрессе: «О MARCе там вообще уже никто не говорит, тем более о UNIMARCе, всё это прошлый век. Теперь только Семантический Web». Самое плохое, что такого рода суждения вполне способны сформировать мнение библиотечного руководства в целом о MARC форматах – и о RUSMARCе в том числе, - как о якобы уже ушедших от нас.

На прошлогодней конференции «ЛИБНЕТ» я говорил в своём докладе следующее: «Колоссальная информация, накопленная в этих форматах, будет существовать и накапливаться ещё долгое время параллельно с более новыми структурами. А значит, и развиваться они должны в необходимой части параллельно с этими структурами». Сейчас так и происходит. И с MARC21, и с UNIMARCом, и с RUSMARCом.

Более того, некоторые способы представления информации в виде Связанных данных, например, через сервер D2R, - а этот, или ему подобные способы грозят стать в ближайшем будущем наиболее популярными, - такие способы вообще не предполагают исчезновения исходных форматов.

Поэтому, давайте не будем спешить с похоронами. Те, кому сейчас 18 лет, возможно и доживут до этого счастливого времени.

На деле посещаемость сайта Национальной Службы развития форматов RUSMARC по-прежнему растёт из года в год. Если в прошлом году средняя посещаемость сайта в рабочие дни составляла 60-75 человек в день – я говорю не о визитах и не о просмотрах страниц, что давало бы более выгодную картину, а именно о посетителях, - то сейчас – 100-125 человек. Это, как мне кажется, уже сравнимо с посещаемостью отдельных библиотек в целом. А ведь сайт Национальной Службы – это узко профессиональный сайт.

Если говорить о географии, то это без преувеличения вся Россия – от Калининграда до Дальнего Востока. Из ближнего зарубежья: Белоруссия, Украина, Литва, Латвия, Молдова, Казахстан, Узбекистан, Азербайджан, Грузия. Из дальнего – Китай, Франция, Польша, Германия, Словения, Болгария, Словакия, Финляндия, США (в частности, округ Колумбия), Канада, Мексика, - это всё осмысленные заходы, судя по количеству визитов и времени пребывания на сайте. Но в целом у нас отметился весь мир, включая Панаму, Бразилию и Аргентину, это при том, что наш сайт – русскоязычный.

По-видимому, на расстоянии RUSMARC видится вполне живым и работоспособным.

Необходимо помнить, что формат RUSMARC создавался с целью экономии государственных средств за счёт устранения дублетной каталогизации, и он по-прежнему служит этой цели. Уж никак не вина формата, в тех случаях, когда повторная каталогизация всё же производится. Иногда это связано с недостаточным охватом библиотек современными корпоративными технологиями, которые просто необходимо далее развивать. А иногда это происходит от неумения или чьего-то нежелания пользоваться инструментами, разработанными совместно ведущими отечественными библиотеками.