УДК 026.06

А. И. Земсков

Новые формы обработки электронных документов

Обзор новейших библиотечно-информаионных технологий, освещённых в докладах на конференции Регионального совета OCLC, в которой принял участие автор публикации.

Ключевые слова: конференция Регионального совета OCLC, связанные данные, электронные документы, библиотечно-информационные технологии, проблемы, перспективы.

26–27 февраля 2013 г. в Страсбурге (Франция) состоялся семинар Регионального совета OCLC *(EMEA Regional Council 2013)*. В его работе (семинар был бесплатным для всех представителей библиотек и родственных организаций) приняли участие около 300 делегатов (50 из них – сотрудники OCLC) из 29 стран.

От России в семинаре участвовали зав. отделом РГБ М. Е. Шварцман и главный специалист ГПНТБ России А. И. Земсков.

Программа 4-й встречи Регионального совета объединена темой «Динамические данные: мир возможностей» (*Dynamic Data: a world of possibilities*). Среди рассматриваемых вопросов – связанные данные (*linked data*), расширенная реальность (*augmented reality*) и глубокая обработка текста (*text mining*). (Материалы заседания можно найти на сайте OCLC: http://www.oclc.org/content/go/en/emearc-mtg-2013.html)

OCLC (*Online Computer Library Center*) – созданная в 1967 г. бесприбыльная организация с регистрируемым членством – одной из своих целей ставит продвижение библиотечных компьютерных технологий. Наряду с ключевой функцией – ведением и поддержанием крупнейшего корпоративного каталога *World Cat* (сейчас в нём свыше 290 млн записей) – OCLC активно проводит собственные исследования по самым передовым библиотечно-информационным технологиям.

Услугами OCLC пользуются более 72 тыс. библиотек из 170 стран. Управляют этой организацией Глобальный совет (*Global Council*), Совет доверенных лиц (*Board of Trustees*) и три региональных совета (*Regional Councils*): Американский – *The OCLC Americas Regional Council*, Азиатский – *The OCLC Asia Pacific Regional Council* и Совет для Европы, Ближнего Востока и Африки – *The OCLC Europe, Middle East and Africa (EMEA) Regional Council*, в котором представлены 7 тыс. организаций из 60 стран, пользующихся услугами OCLC.

Региональные советы проводят ежегодные встречи, где обсуждаются профессиональные проблемы и выбираются делегаты в Глобальный совет.

В составе EMEARC два постоянных комитета – Исполнительный (*Executive Committee*) и Комитет по номинациям (*Nominating Committee*). Председателем EMEARC избрана директор библиотеки Утрехтского университета (*Utrecht University*) Анья Смит (*Anja Smit*).

По масштабам (числу участников и количеству докладов) семинар 2013 года превзошёл все предыдущие: 7 параллельных сессий (на выбор) и обширная программа выступлений (14 докладов); постер-сессия (13 докладов).

Заседания проходили в помещениях Дворца музыки и конгрессов. С ключевыми докладами выступили:сотрудник Гарвардского университета Жан Баптис Мишель (*Jean-Baptiste Michel*), Мари-Кристин Доффе (*Marie-Christine Doffey*) из Национальной библиотеки Швейцарии (*Schweizerische National bibliothek*) и Ричард Уоллис (*Richard Wallis*), главный технолог OCLC (*Technology Evangelist at OCLC*).

Тематика последовательных сессий: «Связанные данные» (*Linked Data*), «Сервисы управления совместным использованием» (*WorldShare Management Services*), «Партнёрство научных библиотек» (*OCLC Research Library Partnership*); «Направления развития метаданных» (*Future Directions for Metadata*).

В первый день работы я присутствовал на семинаре по так называемым связанным данным (СД), который

провёл известный популяризатор и пропагандист этой технологии Ричард Уоллис.

Термином *связанные данные* (*linked data*) обозначают методику публикации в Сети структурированных данных, при использовании которой они взаимосвязаны и, следовательно, станут более полезными. Если традиционное индексирование построено на описании конкретного объекта, то технология СД устанавливает смысловую связь между различными объектами. Как некий грубый пример можно рассмотреть, допустим, простой топор, который в традиционной модели будет описан как «объект 70 см длиной, весом 3 кг, имеющий деревянную рукоятку и металлический наконечник, изготовленный в 1952 г. Колыванским чугунным заводом». Для предлагаемого подхода более существенно указать на связь объекта (топора) с другими объектами (с колуном и рубанком, бревном, поленом, печкой, рубленым домом и т.д.). Как результат, формируется граф, вершины которого – объекты, а ребра – предикаты (сказуемые, заявления).

В этой методике – стандартные технологии, такие как HTTP и URI, но смысл их применения состоит в том, чтобы создавать не только сетевые страницы, пригодные для чтения человеком, но и возможности совместного их использования при автоматическом считывании компьютером. Такой подход позволяет объединять данные из различных источников и осуществлять поиск в нескольких массивах. Связанные данные – это не стандарт, это – набор принципов. В сумме своей СД образуют «сеть в сети», построенную на стандартах семантической сети.

Создатель методики Тим Бернерс-Ли (*Tim Berners-Lee*) впервые использовал термин *связанные данные* в заметках, посвящённых обсуждению проекта семантической сети (*Semantic Web project*). Однако сама идея намного старше и тесно связана с такими концепциями, как модели сети баз данных (*database network models*), перекрестное цитирование в научных статьях и контроль заголовков в библиотечных каталогах (*controlled headings*).

Тим Бернерс-Ли выделил четыре принципа формирования СД:

1. Использование URI для идентификации объектов.

2. Использование HTTP URI таким образом, чтобы на эти объекты можно было ссылаться (*be referred to*) или чтобы их могли просматривать (*dereferenced*) пользователи или специальные пользовательские программы (*user agents*).

3. Предоставление полезной информации об объекте в ходе просмотра с использованием стандартных форматов RDF (*Resource Description Framework*)/XML.

4. Включение в себя ссылок на другие смежные URIs, чтобы облегчить обнаружение иной смежной информации в Сети.

Структурно формат описания ресурсов RDF представляет собой «тройки» – триплеты (*triples*), состоящие из самого объекта + предикат (сказуемое) + субъект.

Выбор предиката достаточно произволен – всё зависит от цели создания массивов СД. Однако крупнейшие поисковые компании – *Google, Bingo, Yahoo, Yandex* – совместно создают кооперативный словарь предикатов.

Универсальный (единый) идентификатор ресурса URI (*uniform resource identifier*) является либо хорошо нам известным универсальным локатором ресурса URL, либо универсальным именем ресурса URN, либо одновременно и тем и другим. Можно сказать, что идентификатор URL – это URI, который, помимо идентификации ресурса, предоставляет ещё и информацию о его местонахождении. А идентификатор URN – это URI, который только идентифицирует ресурс в определённом пространстве имён (и, соответственно, в определённом контексте), но не указывает его местонахождения. Например, URN urn:ISBN:0-395-36341-1 – это URI, который указывает на ресурс (книгу) 0-395-36341-1 в пространстве имён ISBN, но, в отличие от URL, URN не указывает, в каком магазине её можно купить или на каком сайте скачать.

Связанные данные – это подход к представлению данных в машиночитаемом виде, при котором данные могут быть раскрытыми (читаемыми). Иными словами, сетевой идентификатор URI является интегральной неотъемлемой частью показываемых данных, и внешние приложения могут использовать URI для

выполнения различных действий – поиска данных, объединения различных групп данных из множества ресурсов СД и т.п.

Тим Бернерс-Ли в ходе своего выступления на конференции TED–2009 заново сформулировал «чрезвычайно простые», по его мнению, правила СД:

«1. Это все типы объектов, имена которых начинаются с HTTP.

2. Я получаю обратно важную информацию в стандартном формате (это может быть RDF/XML, но допустимы и другие форматы), своего рода полезные данные относительно этого объекта или явления.

3. Я получаю обратно информацию не относительно каких-то формальных признаков (рост, вес или дата рождения). Это информация о их связи с другими объектами».

Примером больших массивов СД может быть программа *s DBPedia*, которая делает содержание *Wikipedia* доступным как RDF. Значение *DBPedia* не только в том, что она обеспечивает доступ к данным *Wikipedia*, а также и ко многим другим массивам данных в Сети, например к *Geonames*. За счёт использования дополнительных ссылок (в терминах RDF-*трипле́тов*) программа может получить дополнительные и, возможно, более точные данные. Ряд других примеров можно найти на сайте *W3C*.

Пример работы OCLC с СД – обработка трёх верхних уровней Десятичной классификации Дьюи (*Dewey Decimal Classification*). Виртуальный авторитетный файл (*Virtual International Authority File – VIAF*) и Фасетные приложения предметной терминологии (*Faceted Application of Subject Terminology– FAST*) опубликованы в виде СД.

В июне 2012 г. OCLC существенно увеличила массив СД, создав возможность перевести в этот формат библиографические записи *WorldCat.org* с помощью вспомогательных наборов разметки *Schema.org mark-up* и специализированных библиотечных программ *Library Extensions* и *Library Vocabulary*. Цель состоит в том, чтобы сделать *World Cat* ценным отправным пунктом для поиска библиографической информации в Сети. В августе 2012 г. опубликован и подготовлен для выгрузки массив связанных библиографических данных, состоящий из 1,2 млн записей *World Cat* (около 80 млн триплетов); объём массива – 860Mb gzip.

Можно полагать, что эти ресурсы будут полезны в качестве источника так называемых сырых СД о произведениях, авторах и издателях. Всё это будет необходимо библиотековедам и тем, кто пожелает проводить культурологические, исторические, социологические или другие исследования на базе богатейшего массива данных, собранного в течение десятков лет для *World Cat*.

Выгрузка осуществляется на основе лицензии *ODC-BY (Open Data Commons Attribution License)* с использованием нормативов (*community norms*), разработанных членами OCLC. Лицензия разрешает использовать данные для исследования и изучения с указанием происхождения набора данных.

СД дают также возможность составлять комментарии, проводить дискуссии в библиотечном сообществе или в более широких кругах.

Возможности, которые открывают СД для мирового библиотечного сообщества, полностью совпадают со стратегической линией OCLC – «совместного с библиотеками построения глобальной системы *Webscal*». С помощью этой методики OCLC поможет библиотекам расширить своё присутствие в Сети, стать более заметными. Продвижение библиотечных коллекций с использованием этого метода – одно из ключевых преимуществ программы *OCLC WorldShare™ Platform*. Программа *OCLC Research* исследует СД с разных точек зрения (в том числе – издателя, потребителя, программиста), работая в партнёрстве со многими организациями, в первую очередь с W3C.

Виртуальный международный авторитетный файл (*Virtual International Authority File – VIAF*) – это интернациональный сервис, предназначенный для удобного доступа к основным мировым авторитетным файлам. Создатели VIAF считают его частью конструкции Семантической сети (*Semantic Web*), позволяющей представлять имена лиц в предпочтительном для пользователей языковом оформлении. Этот проект стартовал как совместная программа Библиотеки Конгресса США (*Library of Congress*), Немецкой национальной библиотеки (*Deutsche National bibliothek*), Национальной библиотеки Франции (*Bibliothèque nationale de France*) и OCLC. За последнее десятилетие к проекту присоединились 20 других национальных

библиотек и агентств из 20 стран.

VIAF сопоставляет и связывает авторитетные файлы национальных библиотек в единую «суперзапись» или «облако записей» для каждого уникального имени, обеспечивает сосуществование национальных и региональных вариантов описания и удовлетворяет потребность в оригинальном описании.

Экспериментальный сервис «Фасетные приложения терминологии предметных рубрик» – FAST (*Faceted Application of Subject Terminology*) – перечислительная схема предметных рубрик, созданная на базе Предметных рубрик Библиотеки Конгресса США (*Library of Congress Subject Headings* – LCSH), – представлена на сайте (http://id.worldcat.org/fast/) и доступна по лицензии *Open Data Commons Attribution License*.

Авторитетный файл FAST, лежащий в основе соответствующих связанных данных, построен в ходе многолетней кооперации *OCLC Research* и Библиотеки Конгресса США. Он в частности позволяет сделать доступным через Сеть богатейший словарь LCSH. Многие географические названия, содержащиеся в LCSH, связаны с географической БД *GeoNames* (http://www.geonames.org/). Более подробную информацию можно найти на http://id.worldcat.org/fast/, а образцы записей – на http://id.worldcat.org/fast/1112076.

Всего к сентябрю 2011 г. в массивах СД насчитывалось около 31 млрд RDF-триплетов, связанных примерно 504 млн RDF-связей (http://en.wikipedia.org/wiki/Linked_data#cite_note-12).

Сотрудники Гарвардского университета – Эрец Либерман Эйден (*Erez Lieberman Aiden*) и Жан Баптис Мишель (*Jean-Baptiste Michel*) – представили в качестве ключевого доклада результаты многоаспектного рассмотрения и изучения миллионов оцифрованных книг, что дало возможность подойти по-новому к проблемам культуры.

Авторы совместно разработали количественный подход к изучению истории и культуры, основанный на компьютерном анализе значительной части исторических записей. Этот подход они назвали *культуромикой* (*culturomics*). Работа привела к созданию веб-сайта *Google Ngram Viewer* (http://books.google.com/ngrams) для просмотра трендов в культуре. В первые 24 часа после открытия сайта его посетило более 1 млн пользователей. Результаты исследований публиковались на страницах и даже обложке журналов «Nature» и «Science» и на первой полосе газет «New York Times», «The Boston Globe», «The Wall Street Journal».

Авторыосновали Гарвардскую обсерваторию культуры (*Harvard's Cultural Observatory*), в которой проводятся междисциплинарные исследования.

В результате работы авторов создан корпус англоязычных электронных текстов за 1800–2000 гг., содержащий примерно 4% всех напечатанных книг. Анализ этого массива позволяет количественно исследовать культурные тренды; самые разнообразные проблемы культуромики, лингвистики, эволюционной грамматики, коллективной памяти, внедрения новых технологий, зарождения и спада известности и каких-либо лиц, и самых разных явлений, событий и т.п.; влияние цензуры (в том числе идентификация субъектов гонения); прогнозировать изменение тех или иных показателей (методом экстраполяции) и т.д.

Пожалуй, можно сказать: если традиционная библиометрия имеет дело с именами авторов, названиями и ссылками в конце статьи, то методика культуромики представляет собой библиометрический анализ полных текстов. Конечно, всё это стало возможным при использовании оцифрованных полных текстов, а их в программе *Google Book* уже более 15 млн.

Приведу один из примеров обработки текстов. Известно, что в английском языке существуют как правильные, так и неправильные глаголы, причём с течением времени число неправильных глаголов снижается. Выбрав 12 учебников по грамматике, учёные проследили частоту использования форм 77 глаголов за период 1800–2000 гг., а также появление в английском языке слов «телефон» (между 1940 г. и 1945 г.) и «радио» (после 1945 г.). Оказалось интересным проследить и за динамикой – появлением и постепенным спадом с течением времени – упоминания тех или иных дат: 1883 г., 1910 г., 1950 г. Вывод – как быстро мы забываем прошлое.

В завершение своего выступления Жан Баптис Мишель ответил на вопросы аудитории:

Какую пользу библиотекам может принести использование вашего подхода?

Сложно предсказать – из-за ограничений, связанных с интеллектуальной собственностью.

Культура – это не только книги; почему же вы говорите о культуре в целом?

Вы правы. Для расширения исследования культуры нужно меньше зависеть от копирайта.

Думаете ли вы, что при анализе книг на другом языке результаты будут иными?

Да, конечно.

Можно ли проводить подобный анализ применительно к науке?

Да, к определённым учёным или научным направлениям и т.д.

В целом семинар был хорошо организован; его участники – достаточно сплочённая группа высококвалифицированных специалистов: они неоднократно встречались на различных мероприятиях OCLC.

Уровень докладов и ход их обсуждения предполагают предварительное ознакомление с тематикой, без чего участие в семинаре окажется малоэффективным.

Сам по себе этот семинар полезен, поскольку без отвлечения на второстепенные вопросы его участники знакомятся с новейшими библиотечно-информационными технологиями, обсуждают связанные с ними перспективы и проблемы.

Очевидным и, по-видимому, неизбежным недостатком является то, что почти все выступления касаются только деятельности OCLC и партнёров этой организации.

Можно полагать, что использование методики анализа полных текстов найдёт применение и в работе российских научно-технических библиотек.