Что такое семантическая цифровая библиотека

© В.А. Серебряков
Вычислительный Центр РАН,
Москва
serebr@ccas.ru

Аннотация

В последние годы в литературе значительное внимание уделяется так называемым «семантическим цифровым библиотекам». Что это такое? В работе на основе анализа проектов и публикаций делается попытка определить такие понятия, как «электронная библиотека», «цифровая библиотека», «семантическая цифровая библиотека».

1 Что такое библиотека

Прежде всего необходимо определить, что такое цифровые библиотеки (в отличие от «электронных» библиотек, под которыми будем понимать программное обеспечение обычных, «книжных» библиотек, (часто называемое АБИС – автоматизированная библиотечная информационна система). Переходя от электронных библиотек к цифровым, можно было бы сказать, что цифровая библиотека – это электронная с цифровым контентом. Это было так в первое время, однако затем контент стал включать живопись, видео и т.д. Так что такое определение устарело.

Электронная библиотека сегодня – это прежде всего формат MARC. Машиночитаемая каталогизация (MARC) – это идея разработки общей системы описания ресурсов библиотек. Она берет начало от работ библиотеки Конгресса еще в 1960-е годы по разработке формата LC MARC для собственных нужд, когда начали использовать компьютеры. MARC-запись стала электронным аналогом бумажного каталога и карточки, который может быть создан в Библиотеке Конгресса и продаваться в библиотеки, которым не придется тратить свои ресурсы для создания почти идентичного набора уже предложенной информации. Даже если библиотека не имеет собственной компьютерной системы, совместимой с MARC, она может приобрести напечатанные на компьютере каталожные карточки, заполненные в соответствии с библиографическими записями в MARC файлах Библиотеки Конгресса. Формат MARC введен в 1987 году, а в 1999 году появился формат MARC21, созданный в результате слияния библиографических форматов США и Канады, и призванный стать «Библиографическим форматом 21 века». MARC 21 является эволюцией исходного LC MARC. Последующие издания были опубликованы в 1990 году, 1994 и 2000 года. MARC21 поддерживается библиотекой конгресса США, и используется в основном в США и Великобритании. В настоящее время существуют две группы, ответственные за рассмотрение и пересмотр формата MARC 21: Комитет Marbi (машиночитаемая библиографическая информация) и Консультативный комитет MARC. Каждый год появляется новое официально опубликованное издание MARC 21 в Интернете с учетом изменений в библиотечной среде.

В 1977 году был выпущен формат UNIMARC, который был призван стать посредником между любыми национальными стандартами MARC. Формат UNIMARC включает поля, необходимые для описания монографий, сериальных изданий, нотных записей, видео, изображений и прочих документов. Эти поля делятся на общие, использующиеся при описании любого вида документа, и специфические, используемые только для описания их определенных видов. Этот формат поддерживается международной организацией IFLA, и используется в основном в Европе и Азии.

Программное обеспечение обычной, «книжной» библиотеки состоит из нескольких базовых компонент, которые можно разделить на два блока: блок работы с читателями, включающий проверку обслуживание прав читателя, выдачу, прием и заказ книг и т.д., и блок обслуживания фонда, куда условно можно отнести заказ и покупку литературы, списание, постановку на учет (включая подготовку библиографических записей) и т.д.

1.1 Что такое цифровая библиотека

Самое простое определение, которое можно дать – «Цифровая библиотека – это электронная библиотека с цифровым контентом». Более неопределенно можно сказать, что ЦБ – это информационная система, основным назначением которой является доступ к цифровым материалам. Здесь подчеркивается, что организация самой информационной системы может быть произвольной, важно, что вся эта организация нацелена на доступ к цифровому контенту (тексты, видео, аудио и т.д.). Wikipedia определяет цифровую библиотеку так: «Цифровая библиотека – это библиотека, в которой коллекции хранятся в

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

цифровых форматах (в отличие от печатного, микроформата или другого носителя) и собираются с помощью компьютеров».

Еще одним предшественником ЦБ были библиотеки программ. Изначально они были предназначены для размещения и использования объектов операционных систем: библиотеки для связывания объектного кода, библиотеки исходного кода, компилированного объектного кода для повторного использования. Они возникли из потребностей ОС, чтобы находить и загружать компоненты, и того факта, что существующие файловые системы не обеспечивали работу в реальном времени. Возникшая структура остается по-существу и сегодня; за справочником элементов библиотеки, который дает имена и другие метаданные содержащихся объектов, следуют в том же наборе данных или файле двоичные данные для каждого объекта, на который ссылается элементы каталогов.

Часто в связи с ЦБ используется термин «коллекция», под которым имеют в виду определенным образом организованный набор как правило однородных цифровых объектов.

«Основой цифровой библиотеки является коллекция цифровых объектов, которые представляют интерес как таковые (в первую очередь для чтения, прослушивания, просмотра людьми, но и для использования программами), а не просто указания на другие объекты. Примеры:

- Коллекция оцифрованных книг (в отличие от просто интернет-каталога),
- Коллекция биографий (в отличие от базы данных персонала),
- Коллекция устных историй,
- Набор программных модулей (многие так и рассматривают DL)» [5].

Рассмотрим теперь определения цифровых библиотек, приводимые различными авторами и их критику.

ru.wikipedia.org дает такое определение.

Электро́нная библиоте́ка – упорядоченная коллекция разнородных электронных документов (в том числе книг, журналов), снабженных средствами навигации и поиска. Может быть веб-сайтом, где постепенно накапливаются различные тексты (чаще литературные, но также научные и любые другие, вплоть до компьютерных программ) и медиафайлы, каждый из которых самодостаточен и в любой момент может быть востребован читателем. Электронные библиотеки могут быть универсальными, стремящимися к наиболее широкому выбору материала (как Библиотека Библиотека Максима Мошкова или Либрусек), и более специализированными, как Фундаментальная электронная библиотека или проект Сетевая Словесность.

Возникает ряд вопросов. Что значит упорядоченная коллекция? Что значит разнородных? Средствами навигации снабжены документы? Только навигации и поиска, т.е. никаких связей нет. Что значит «Может быть веб-сайтом»? А может не быть? Т.е. может быть не привязана к Интернет. «Постепенно накапливаются», а если не постепенно? Почему чаще литературные? Самодостаточен, т.е. в некотором роде отсутствуют связи между ресурсами. Библиотека Мошкова и Фундаментальная электронная библиотека радикально отличаются друг от друга: первая просто набор файлов, вторая пронизана ссылками (HTML).

«Под электронными библиотеками понимаются информационные системы, которые автоматизируют решение основных проблем организации работы с документами» [7].

В соответствии с таким определенем наилучшей электронной библиотекой является система документооборота.

В [4] дается следующее определение ЦБ (в оригинале «Электронной библиотеки»).

«Электронные библиотеки – это организации, в том числе специализированный персонал, представляющие доступ читателей к электронным ресурсам. Кроме того они выполняют отбор, структурирование, предоставление интеллектуального доступа, интерпретацию, распространение, сохранение целостности и обеспечение сохранности в течение длительного времени наборов электронных документов для удобного доступа к ним определенным сообществам специалистов.

В соответствии с данным определением основными компонентами ЭБ являются: специалисты, информационные ресурсы (документы) и информационные технологии.

Электронные библиотеки реализуют набор функций для обеспечения читателям полного доступа к множеству распределенных и разнородных документов, содержащих информацию и знания, интегрируя их в единое информационное пространство».

1.2 Цифровые библиотеки или информационные системы?

С другой стороны, ясно, что цифровую библиотеку можно считать информационной системой. А почему бы не считать любую информационную систему цифровой библиотекой? Любая информационная система в конце концов имеет дело с цифромым контентом. Есть ли все-таки разделительная линия, выделяющая цифровые библиотеки из общего класса информационных систем?

В [1, 2, 4] описаны некоторые проблемы ЭБ, основными из которых являются следующие:

- Проблема интеграции разнородной информации (электронных ресурсов,

пользовательских профилей, таксономий) на основе различных метаданных, содержащих выразительные семантические описания.

- Проблема поддержки взаимодействия с другими информационными системами (и не толькоЭБ) либо с помощью метаданных, либо на уровне коммуникации или с помощью обеих возможностей. При этом в качестве единого языка взаимодействия между системами может использоватьсяязык RDF (Resource Description Framework).

- Проблема обеспечения надежного, удобного и адаптируемого поиска и интерфейсов просмотра электронных документов, усиленных работой с семантикой [7].

«ЦБ можно охарактеризовать диапазоном целей, которым она служит, или областью в которой он работает, например, обучение, образование, электронное правительство, электронная коммерция (B2B или B2C), развлечения, и более специфические цели, такие как обеспечение информации, связанной с работой, поддержка домашних заданий студентов, поддерживая внутренней работы организации, поддержка клиентов организации, поддержка связи между пользователями и т.д.» [5].

«1. ЦБ имеет много функций и должна интегрировать поддержку информационного поиска, задачи пользовательской работы, производство информации и сотрудничество.

2. ЦБ связывает многие виды информационных объектов в различных форматах (в том числе документы и базы данных) во всех средствах массовой информации в сложную структуру» [5].

Рассмотрим еще несколько определений ЦБ в контекте информационных систем.

«Термин Цифровая библиотека (ЦБ) используется для диапазона систем, от цифрового объекта и хранилищ метаданных, системы ссылка-связь, архивов и систем управления контентом до сложных систем, которые объединяют в себе передовые цифровые библиотечные услуги и поддержку научных исследований и практических сообществ» [5].

Ничего специфичного для ЦБ в этих определениях нет, это все также относится и к информационным системам.

Рассмотрим, как некоторые авторы определяют функции ЦБ [5].

«Цифровые библиотеки сталкиваются со многими проблемами, в том числе:

- Поиск текста, изображения, звука и составных объектах мультимедиа.
- Семантически улучшенный поиск для извлечения из свободного текста и изображения и лучшего использования проставленных пользователем меток.
- Многоязычный поиск.
- Поиск во многих системах синтаксического и семантического взаимодействия.
- Нахождение ответов, а не только документов; рассуждения и логический вывод».
- Интеграция многих форматов сохранения.
- Интеграция библиотек, архивов, музеев а также баз данных и других информационных систем.
- Интеграция чтение / просмотр / прослушивание, доступ к базе данных, обработка данных и создание.
- Интеграция издательских и коммуникационных платформ».
- Сервисы Распространение и уведомления. Современные цифровые библиотеки должны помочь своим пользователям в доступе к метаданным в различных форматах, позволяющих, среди других, построения мэшапы сервисов и контента.
- Сервисы безопасности и политики Assurance. Библиотека должна приспосабливаться к различным усилениям политики; она должна обеспечить гибкие механизмы аутентификации и контроля доступа.
- Сервисы сохранения. Цифровая библиотека должна обеспечить управление версиями, архивирования (резервного копирования и восстановления) а также, отслеживания происхождения (особенно в контексте открытого мирового подхода семантических и социальных технологий), и отслеживание истории событий, связанных с информационными объектами. Должно быть обеспечено, что отношения между объектами и дополненая информация поддерживаются сервисами сохранения.
- Сервисы обеспечения качества. Особое внимание следует уделять качеству сервисов на основе метаданных; семантическая цифровая библиотека должны обеспечить эффективность, безопасность и семантику поддержки метаданных. Эффективность может быть достигнута, например, путем жесткого кодирования части метаданных; ограничений на действия, которые могут быть выполнены над метаданными, могут повысить уровень безопасности. Семантика метаданных можно определить через значения новых концепций».

Из вышеприведенного можно видеть, что при таких определениях любую информационную систему можно рассматривать как цифровую библиотеку.

2. Что такое семантическая цифровая библиотека

Само по себе слово «семантический» означает не более, чем «смысловой», т.е. в отрыве от контекста не означает ничего. Этот термин (когда-то используемый в теории языков программирования) стал активно употребляться в контексте «семантический WEB» в противовес

«несемантическому WEB», основанному на гиперссылках. Фактически сегодня под «семантической моделью WEB» имеется в виду использование RDF модели для представления информции. Но что такое RDF модель? Это всего навсего использование бинарных отношений, т.е. связей, между объектами и соответствующие словари RDF, обобщающие и стандартизующие их использование. Это внесло колоссальный прорыв в технологии WEB. Но в конце концов, практически все данные, в частности, конечно, и данные цифровых библиотек, хранятся сегодня в реляционных базах данных, также представляющих собой отношения, только вообще говоря, многоместные.

Термин «семантический» не вносит ничего нового в технологии цифровых библиотек. Единственное, что может быть тут стоит отметить, что в обычных цифровых библиотеках эти связи между объектами используются недостаточно активно, хотя в рамках формата MARC, разработанного Библиотекой Конгресса США, предусмотрены так называемые «авторитетные» файлы, хранящие инофрмацию о персонах и организациях. Но эти данные недостаточно формализованы, чтобы их легко можно было использовать для установления всех необходимых связей.

Поэтому термин «Цифровые семантические библиотеки» осмысленно употреблять только в контексте WEB, а именно имея в виду интеграцию цифровых библиотек в контекст семантического WEB». А это означает:

- Разработку стандартов обмена RDF информацией. В качестве примера можно привести онтологии MADS и MODS, разработанные Библиотекой Когресса США для авторитетных файлов и библиографических записей.
- «семантическую» интеграцию библиотек между собой, т.е. возможность, способность цифровых библиотек обмениваться такой информацией.
- «погружение» цифровых библиотек в семантический WEB, т.е. интеграцию с другими, небиблиотечными данными, например, с соцсетями.
- Взаимодействие с данными из Linked Open Data (LOD), например, извлечение данных из LOD в библиотеку и наоборот, публикация собственных данных в LOD.

«Включение семантических данных и обработки в DL предполагает использование метаданных объектов в такой библиотеке и обеспечение доступа пользователей к семантически более мощным поисковым системам. Метаданные, как правило, выражается в Синтаксисе RDF» [3].

Интересно отметить еще одно обстоятельство. В контексте цифровых семантических библиотек часто упоминают онтологии. Насколько это важно и характерно именно для цифровых семантических библиотек? Опять происходит некая подмена. Онтологии в современном понимании могут использоваться в трех целях: 1) как модель данных более высокого уровня по сравнению с использовавшимися моделями раньше, а именно моделью «сущность-связь» и объектной; 2) для поддержки интеграции данных в пространстве Интернет и 3) для реализации восзможности осуществления логического вывода. Для реализации 1-й цели в ЦСБ онтологии используются в той же мере, в какой они используются для разработки информационных системных в прикладных областях. Для реализации 2-й цели отологии активно используются в той же мере, в какой они используются для интеграции данных в Интернет. Для 3-й цели в приложении к ЦСБ онтологии не используются.

В контексте ЦБ упоминаются соц. сети, обучающие системы, архивные системы и их связь с пользователями. Все это было и не называлось ЦБ.

«Основной целью семантической цифровой библиотеки является предоставление нахождения информации превосходящее решения, обеспечиваемые текущими цифровыми библиотеками. Пользователи должны иметь возможность использовать взаимосвязанную информацию о ресурсах в процессе просмотра, фильтрацию или нахождение подобных информационных объектов. Средства уточнения запроса должны адаптировать свои результаты для решений, соответствующим пользовательским профайлам; средства должны использовать сложные семантические отношения между результатами. Наконец, семантическая цифровая библиотека должна предлагать различные рекомендательные сервисы, например, на основе контекста и ресурса (ресурсов) или аннотации на основе совместной фильтрации. Поисковая система должна позволять использовать информацию о различных типах носителей, сложных объектах, потокового и пространственно-временных ресурсах. В случае ресурсов со сложными аннотациями важно поддерживать поиск на основе содержимого вместе с алгоритмами поиска, основанными на сходстве. В случае гетерогенных конкурентных сетей контент-провайдеров, семантическая цифровая библиотека должна осуществлять алгоритмы запроса, основанные на торговле, для поддержки пользователей в их поиске» [3].

«Одной из наиболее отличительных особенностей семантических цифровых библиотек является дополнительное пополнение аннотаций исходной информации, представляемые в ходе процесса загрузки ресурса. Ожидается, что семантические цифровые библиотеки могут обеспечить как автоматизированные, так и пользовательские аннотации. Последние должны использовать силу социальных сетей, то есть аннотации сообщества, пометки, и рейтинг» [3].

Заключение

Резюмируя вышеприведенный краткий обзор, можно остановиться на следующих определениях.

Электронная Библиотека (ЭБ, АБИС) – средство автоматизации работы обычных, «книжных» библиотек, основанное как правило на технологиях MARC.

Цифровая Библиотека (ЦБ) – информационная система, ориентированная на действия (поиск, доступ и т.д.) с цифровым контентом (тексты, аудио, видео и т.д.). В этом смысле ЦБ може быть, а может и не быть ЭБ.

Семантическая Цифровая Библиотека (СЦБ) – ЦБ, ориентированная на интеграцию в Semantic Web.

Литература

[1] Ding Hao. A semantic search framework in peer-to-peer based digital libraries. – NTNU, Norway, 2006.

[2] Sebastian Ryszard Kruk, Adam Westerki, and Ewelina Kruk. Architecture of Semantic Digital // Semantic Digital Libraries / Editors: Sebastian Ryszard Kruk, Bill McDaniel. – Springer, 2009.

[3] Sebastian Ryszard Kruk and Bill McDaniel. Goals of Semantic Digital Libraries // Semantic Digital Libraries / Editors: Sebastian Ryszard Kruk, Bill McDaniel. – Springer, 2009.

[4] A.A. Shiri. Digital library research: current developments and trends // Library Review. – 2003. –Vol. 52. – P. 198–202.

[5] Dagobert Soergel. Digital Libraries and Knowledge Organization // Semantic Digital Libraries / Editors: Sebastian Ryszard Kruk, Bill McDaniel. – Springer, 2009.

[6] Sukhdev Singh. Digital Library: Definition to Implementation [Электронный ресурс]. – http://arizona.openrepository.com/arizona/bitstream/ 10150/106534/1/ lecture_rcc_26jul03.pdf

[7] Ле Хоай, А.Ф. Тузовский, Разработка семантических электронных библиотек. Доклады ТУСУРа, № 2 (24), часть 2, декабрь 2011.

Semantic digital libraries. What is it?

Vladimir Serebryakov

In recent years, considerable attention is paid to the so-called “semantic digital libraries”. What is it? In this paper, based on analysis of projects and publications an attempt is made to define concepts such as “electronic library”, “digital library”, “semantic digital library”.