Функциональные требования к модели электронной библиотеки по научному наследию*

О. А. Федотова
Государственная публичная научная библиотека СО РАН
o4f8@mail.ru

Доклад посвящен обсуждению функциональных требований к модели электронной библиотеки (ЭБ) по научному наследию и архитектуре ее организации. Эти требования определяются, во-первых, информационными потребностями исследователей, а во-вторых, обеспечением надежного и долговременного хранения информации.

Ключевые слова: научное наследие, поиск информации, электронные библиотеки, библиографические базы данных, распределенные информационные ресурсы.

1. Введение

Актуальность настоящей работы обусловлена тенденцией к переводу разнородной информации с бумажных носителей в цифровую форму, к созданию крупномасштабных информационных хранилищ и организации разграниченного пользовательского доступа к этой информации. При этом главным является обеспечение интероперабельности создаваемых информационных систем (ИС), что приводит к требованию соответствия всех интерфейсов и протоколов соответствующим мировым стандартам [1, 2].

Научное наследие — это опубликованные результаты научных исследований и экспериментов, библиографические и фактографические базы данных, сведения об ученых, их научной деятельности, публикациях, проектах и т. п., а также большое количество неопубликованных документов, таких как отчеты, письма, воспоминания, записки, фотоматериалы и т. п. Эти ресурсы представляют интерес для научного сообщества и представителей деловой общественности.

Однако в настоящий момент значительная часть информационных ресурсов по научному наследию недоступна широкому кругу научной общественности, а ресурсы, представленные в Интернет, существенно разрознены, недостаточно систематизированы и структурированы. При создании их описаний недостаточное внимание уделяется вопросам интероперабельности — слабо применяются соглашения и рекомендации по стандартизации представления документов и средства интеграции разнородных информационных ресурсов.

Для документов по научному наследию важной проблемой является идентификация ресурсов, определяющая конкретно для каждого факта, кто является его автором, где и когда он получен, с какими другими фактами он связан. Для этого необходима поддержка различных уровней абстракции при описании информации от кратких описаний до очень подробных описаний информационных объектов. Для поддержки

*Работа выполнена при финансовой поддержке РФФИ (проекты 12-07-00472а, 11-07-00561а, 13-07-00258а), а также в рамках программы Государственной поддержки научных школ РФ (грант НШ-6293.2012.9).

сложных функций поиска и классификации информации недостаточно хранить только полнотекстовые описания. Необходимы поддержка поиска по атрибутам, полнотекстового поиска, а также просмотр ресурсов по категориям и словарям-классификаторам.

Для наиболее полного удовлетворения потребностей научного сообщества необходимо создание интеллектуальных ИС, в качестве составных компонентов которых выступают, наряду с традиционной ИС, еще и рассуждающая ИС, формализующая правила логического вывода, а также интеллектуальный интерфейс.

В большинстве существующих ЭБ документы являются слабо структурированными: хотя и снабженными метаданными, но содержащими неструктурированные элементы. Поэтому, актуальной задачей является разработка теоретических основ и моделей создания ИС, способных в автоматизированном режиме извлекать метаданные из электронных документов достаточно произвольной структуры. Ее решение позволит получать новую информацию и знания.

Различные аспекты поиска информации и методология разработки информационных систем обеспечения различных аспектов научной деятельности предложены в публикациях Федотова А. М. и Барахнина В. Б. [3].

Современными технологиями в информационном обеспечении научных исследований активно занимается группа под руководством Серебрякова В. А. и Бездушного А. Н. (проект создания «ЕНИП» [4]). На основе разработанных технологий ими была создана ЭБ «Научное наследие России»[1].

Следует также упомянуть систему Current Research Information System (CRIS[2]), предназначенную для предоставления доступа к исследованиям и распространению научной информации. Для поддержки этой системы разработан международный стандарт хранения научных результатов CERIF[3].

2. Определение электронной библиотеки

Проблема поиска информации — одна из вечных проблем человечества. Чтобы решить проблему доступа к информации человечество создало библиотеки — универсальную систему хранения, систематизации и каталогизации «информации и знаний» [3].

Электронная библиотека (ЭБ) — это структурированная каталогизированная коллекция разнородных электронных документов, снабженная средствами навигации и поиска (в отличие от печатных изданий, микрофильмов и других носителей). ЭБ способна не только обеспечить многосторонний поиск в каталоге, но и предоставить пользователю непосредственно найденный ресурс (публикацию, документ, фотографию, описание факта и др.), а также дополнительные сведения о нем, например, информацию об авторах, библиографию, организации и т. п.

За высокой популярностью слов «электронная библиотека» стоит не только и не столько дань моде, сколько попытка охарактеризовать новый феномен — возникновение принципиально нового класса систем, призванных аккумулировать и распространять информацию в электронной форме. А большой интерес к самим системам данного класса объясняются потребностями общества и наличием развивающихся возможностей по их удовлетворению. В связи с этим можно сформулировать основные цели, стоящие перед ЭБ:

[1] http://e-heritage.ru/index.html

[2] http://www.eurocris.org/

[3] CERIF — Common European Research Information Format.

– обеспечение доступа к информации;
– сохранение научного и культурного наследия;
– повышение эффективности научных исследований и обучения.

В существующих разработках ЭБ, как правило, поиск и доступ к информации обеспечивается только посредством визуальных графических интерфейсов. Это хорошо для пользователя-человека, но не годится для пользователя-системы. Для обеспечения функций поиска вне графических интерфейсов требуется поддержка специальных сетевых сервисов и языков запросов. В идеальном случае все ИС должны поддерживать единый поисковый профиль и единый язык запросов.

Однако в общем случае под словосочетанием «электронная библиотека» могут фигурировать совершенно различные объекты, такие как архивы цифрового контента и наборы программного обеспечения для управления этим контентом. Электронной библиотекой может называться система сетевых сервисов, предоставляющих доступ к цифровому контенту, объединенных единой системой управления этим доступом [5]. Такое определение ЭБ полностью соответствует определению традиционной библиотеки как организации в системе, например, министерства культуры [1].

В настоящее время нет какой-либо универсальной системы поддержки ЭБ, которая отвечала бы всем требованиям и ожиданиям пользователей. Анализ существующих систем ЭБ (см., например, [6]) показывает их разнородность на нескольких уровнях:

– на уровне информационной модели, которую они обеспечивают;
– на уровне поддержки пользователей и групп пользователей;
– на уровне функциональных возможностей.

Из-за этой разнородности и игнорирования нужд пользователей возникает ряд проблем:

– интеграция информации из различных ЭБ;
– сравнение ЭБ по предоставляемой функциональности;
– оценка и сравнение производительности различных систем ЭБ;
– добавление новых типов хранимых объектов;
– добавление новых функциональных возможностей;
– резервное копирование.

В настоящее время существуют достаточно мощные ИС, удовлетворяющие в той или иной степени потребности научных работников в информации, однако основной недостаток большинства систем — ограниченность возможностей обеспечения интеграции ресурсов как внутри каждой из систем, так и с внешними системами. Основу разработки ЭБ составляют стандарты и международные рекомендации, формирующие профиль ЭБ, под которым понимается набор из одного или нескольких базовых нормативно-технических документов (стандартов и спецификаций), ориентированных на решение определенной задачи (реализацию заданной функции либо группы функций приложения или среды) с указанием, при необходимости, выбранных классов, подмножеств, опций базовых стандартов, которые являются необходимыми для выполнения конкретной функции [7]. Наиболее важным является профиль метаданных информации, циркулирующей в системе. Выбор профиля должен основываться на выполнении следующих требований:

– включать в себя основные типы информации, требующейся для поддержки научной работы;

– быть открытым, т. е. обеспечивать доступ к соответствующей информации по этим описаниям;
– быть расширяемым, т. е. обеспечивать возможность детализации описаний;
– обеспечивать возможности интеграции информации;
– обеспечивать возможности уникальной идентификации информации;
– обеспечивать возможности размещения и поиска информации в распределенной среде;
– быть ориентированным на современные и перспективные технологии описания и использования информации;
– обеспечивать возможности интероперабельности с внешней средой.

При работе с цифровыми объектами человечество уже выработало определенный набор стереотипов, отсутствие которых вызывает дискомфорт [1]. Одним из элементов этого набора является требование наличия взаимных ссылок между цифровыми объектами, проявляющихся, например, в виде гиперсвязей в пользовательских графических интерфейсах просмотра информации. Реализация взаимных ссылок в цифровых документах не представляет большой сложности, однако при этом проявляются специфические моменты. Во-первых, электронный объект с реализованными связями уже не совсем соответствует своему печатному оригиналу. Во-вторых, внедренные в объект связи должны быть гарантировано актуальными. Так появляется требование обеспечения ссылочной целостности данных. Это очень жесткое требование, которое трудно обеспечить даже в хорошо формализованных системах управления базами данных. Результат — новый цифровой объект как самосогласованное хранилище цифрового контента, или база данных цифровых объектов.

С другой стороны, в ЭБ объекты хранения могут содержать информацию, которая не имеет к объектам хранения традиционных библиотек вообще никакого отношения. Речь может идти об электронных копиях элементов хранения традиционных архивов, о видео-, аудио- информации, полученной разными способами, о научных или других фактах и т. п.

Для решения возникающих проблем создаются концептуальные модели, обобщающие накопленный опыт в сфере создания и использования ЭБ.

3. Функциональные требования к модели электронной библиотеки по научному наследию

Как уже отмечалось выше, основными целями создания ЭБ по научному наследию являются:

– предоставление научным работникам быстрого доступа к информационным ресурсам по научному наследию;
– предоставление результатов фундаментальных научных исследований мировому сообществу;
– предотвращение утраты ценных научных коллекций для будущих поколений ученых;
– создание новых технологий научных исследований, эффективного инструментария для их проведения.

Как известно большая часть научной информации быстро устаревает. Но это не относится к материалам по научному наследию. Для этого типа информационных ресурсов важно хранить описание жизненного цикла этих ресурсов и иметь возможность восстановить состояние ресурса на любой момент времени. Кроме того, существуют информационные ресурсы, которые могут быть доступны длительное время. К таковым, например, относятся документы, имеющие длительную юридическую силу, патенты, мультимедийная информация об исторических событиях, которая может быть востребована через любой период времени. Кроме того, научные отчеты институтов, речи ученых, письма и служебные записки могут также иметь огромную историческую значимость, становясь более ценной со временем. Поэтому ЭБ должна поддерживать возможность длительного хранения информационных ресурсов с возможностью восстановления их.

Для документов по научному наследию важной проблемой является идентификация информационных ресурсов [8, 9], определяющая конкретно для каждого факта, кто является его автором, где и когда он получен, с какими другими фактами он связан. Для этого необходима поддержка различных уровней абстракции при описании информации от кратких описаний, до очень подробных описаний информационных объектов.

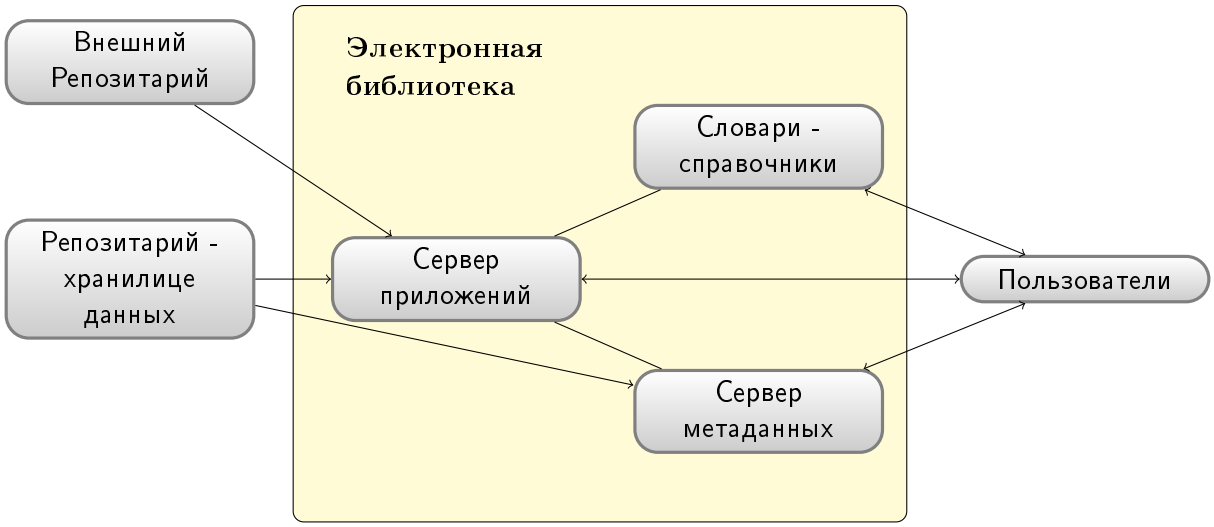

Рис. 1. Архитектура электронной библиотеки

Исходя из целей ЭБ по научному наследию и анализа существующих систем, направленных на поддержку научных исследований, можно сформулировать следующие функциональные требования к модели ЭБ по научному наследию:

- надежное долговременное и защищенное от исчезновения хранение информации;
- актуальность, полнота, достоверность происхождения документов;
- историчность информации;
- географическая привязка информации;
- наличие большого числа словарей-классификаторов (справочников), для обеспечения идентификации и классификации ресурсов;
- поддержка неоднородных и слабо структурированных информационных ресурсов;
- поддержка взаимосвязей информационных ресурсов;
- предоставление информации пользователю в виде, выбранном пользователем;

– наличие интеллектуальных служб обслуживания запросов пользователя;
– наличие программных интерфейсов для поддержки аналитической работы пользователя с помощью программных приложений;
– поддержка требований интероперабельности как на программном, так и на семантическом уровне;
– поддержка работы с внешними источниками.

Наиболее важным выводом из вышесказанного является то, что информационная модель ЭБ должна быть многоуровневой и состоять как минимум из следующих компонент [10, 11]: хранилище данных — репозиторий, сервер метаданных, сервер приложений, словари-справочники (см. рис. 1).

4. Выбор метаданных для ЭБ по научному наследию

Ввиду того, что информация в ИС является отображением реальных или материальных сущностей (предметов, процессов, явлений, персон, публикаций и т. п.), следует рассматривать ИС как множество информационных объектов — наборов данных, представляющих (описывающих) эти сущности в ИС. В работах [2, 8] был определен профиль ЭБ как необходимый набор стандартов и компонент ИС.

Эффективным средством описания информационных объектов в ИС являются метаданные — данные, являющиеся неотъемлемой частью информационного объекта и описывающие реальный объект или группу объектов (см. рис. 2).

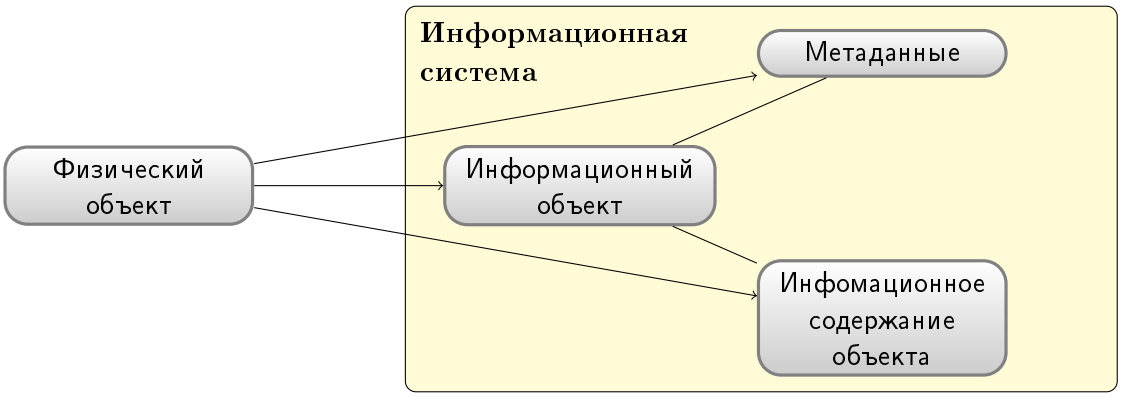

Рис. 2. Структура информационной системы

В настоящее время существует большое количество систем метаданных, предназначенных для описания различных классов информационных объектов. Использование систем метаданных (схем данных) пока еще недостаточно формализовано. Информационные системы, ориентированные на одинаковые классы информационных объектов, используют различные, часто оригинальные схемы и форматы метаданных, а также разные подходы к решению прикладных задач. Решением этих проблем занимаются многие организации во всем мире, например, W3C, DCMI, OCLC, IFLA, IETF, ISO.

Под интероперабельностью любой ИС, в том числе и ЭБ, понимается степень ее способности взаимодействовать с другими ИС, в том числе и с человеком. Но если при взаимодействии с человеком (как с информационной системой) основная нагрузка на обеспечение взаимопонимания ложится на человека, который в состоянии обработать

даже плохо организованную информацию, то для обеспечения эффективного взаимодействия между собственно информационными системами требуются специальные технологические методы и общие соглашения.

Метаданные необходимы для решения следующих задач:

- предоставление сведений об объекте для получения представления о его содержании, структуре, способах использования и т. д.;
- сбор и систематизация информации об объектах описания;
- выбор из множества объектов определенного подмножества по формальным признакам и сопоставление объектов по формальным признакам;
- внутрисистемные технологические задачи, связанные с обеспечением подготовки объектов, размещением объектов в информационном фонде и т. п.;
- внешние технологические задачи, связанные, прежде всего, с обменом данными с внешними информационными системами.

Для формирования простых метаданных применяются несколько стандартов, являющихся расширениями рекомендаций Dublin Core[4]. Используемый профиль определяет список элементов данных (полей), необходимых для создания записи соответствующего типа и раскрывает содержание элементов данных [1, 2]. Для эффективной работы сервера приложений необходимо использовать набор словарей-классификаторов, содержащих как классификационные признаки, так и наборы ключевых терминов (с отношениями порядка), по которым производится систематизация и классификация материала.

Словари (ключевые признаки) — это особый вид метаданных, которые отражают наиболее существенные свойства объекта, имеющие наибольшее значение с точки зрения ИС, и их специфика определятся терминологией конкретной предметной области, которой посвящена ЭБ.

Имеется ряд российских (например, УДК, ГРНТИ) и международных (например, MSC-2000[5], ORTELIUS[6]) словарей для классификации научных данных. Однако в целом эти словари содержат только общенаучную информацию и не годятся для систематизации материалов по научному наследию конкретной научной школы. Необходимо рассматривать различные типы ключевых терминов, а именно:

- ключевые термины в стандартном понимании;
- ключевые термины, описывающие персону;
- ключевые термины, описывающие организацию;
- ключевые термины, описывающие временные периоды;
- ключевые термины, описывающие географические понятия.

Метаданные существенным образом зависят от природы и структуры объектов реального мира, от способа представления их в виде информационных объектов и от специфики ИС. Учитывая это, необходимо классифицировать описываемые объекты. Законченная совокупность правил, достаточная для формирования метаданных в определенном классе ИС и/или для решения определенного класса задач над информационными объектами представляет собой систему метаданных.

Функционирование ИС связано с разнообразными процессами по созданию метаданных, их модификации, проверке корректности, предоставлению метаданных конечному

[4]Dublin Core Metadata Initiative — `http://www.dublincore.org/`

[5]`http://www.ams.org/msc/`

[6]`ftp://ftp.cordis.lu/pub/cerif/docs/ortelius.doc`

пользователю и решению прикладных задач. Все эти процессы являются взаимосвязанными, их выполнение усложняется, как правило, большим количеством объектов, на представление и работу с которыми нацелена ИС. Реализация этих процессов и управление ими требуют специальных средств и методов, которые в совокупности с метаданными можно рассматривать как отдельную подсистему — систему метаинформационного сопровождения.

Список литературы

[1] *Жижимов О. Л., Мазов Н. А., Федотов А. М.* Некоторые заметки об эволюции цифровых репозитариев традиционных библиотек к полнофункциональным электронным библиотекам // Вестник Владивостокского государственного университета экономики и сервиса. Территория новых возможностей. – 2010. Т. 7, № 3. С. 55–63.

[2] *Федотов А. М., Барахнин В. Б., Жижимов О. Л., Федотова О. А.* Технология создания корпоративных информационных систем учета трудов научных работников // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2011. — т. 9. вып. 2. С. 31–41.

[3] *Шокин Ю. И., Федотов А. М., Барахнин В. Б.* Проблемы поиска информации. Новосибирск: Наука, 2010. 198 с.

[4] *Бездушный А. Н.* Интеграция метаданных Единого Научного Информационного Пространства РАН / Бездушный А. Н., Бездушный А. А., Серебряков В. А., Филиппов В. И. — М.:ВЦ РАН, 2006.

[5] *Шокин Ю. И., Федотов А. М. Жижимов О. Л., Гуськов А. Е., Столяров С. В.* Электронные библиотеки - путь интеграции информационных ресурсов Сибирского отделения РАН // Вестник КазНУ, специальный выпуск. — г. Алматы, Казахстан, Казахский национальный университет им. аль-Фараби. 2005. № 2. С. 115–127.

[6] *Candela L., Castelli D., Fuhr N., Ioannidis Y., Klas C.-P., Pagano P., Ross S., Saidis C., Schek H.-J., Schuldt H., Springmann M.* Current Digital Library Systems: User Requirements vs Provided Functionality. IST-2002-2.3.1.12. Technology-enhanced Learning and Access to Cultural Heritage. March 2006.

[7] ГОСТ Р ИСО / МЭК ТО 10000-2-99. Информационная технология. Основы и таксономия функциональных стандартов. Часть 2. Принципы и таксономия профилей ВОС.

[8] *Федотов А. М., Барахнин В. Б., Жижимов О. Л., Федотова О. А.* Проблемы создания информационных систем учета трудов научных сотрудников СО РАН // Труды IV Междунар. конф. «Системный анализ и информационные технологии» (САИТ-2011) (Абзаково, Россия, 17–23 августа 2011). — Челябинск: ЧелГУ, 2011. С. 85–91.

[9] *Федотов А. М., Жижимов О. Л., Князева А. А., Колобов О. С., Мазов Н. А., Турчановский И. Ю., Федотова О. А.* Проблемы авторитетного контроля для распределенных электронных библиотек и библиографических баз данных // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2011. — Т. 9. Вып. 1. С. 89–101.

[10] *Федотов А. М.* Методологии построения распределенных систем // Вычислительные технологии. 2006. Т. 11. С. 3–17.

[11] *Жижимов О. Л., Федотов А. М., Федотова О. А.* Построение типовой модели информационной системы для работы с документами по научному наследию // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2012. — Т. 10, № 3. С. 5–14.