УДК 004.9

А. М. Федотов, **В. Б. Барахнин**, **О. Л. Жижимов**, **О. А. Федотова**

Институт вычислительных технологий СО РАН
пр. Лаврентьева, 6, Новосибирск, 630090, Россия

Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия

Государственная публичная научная библиотека СО РАН
ул. Восход, 15, Новосибирск, 630200, Россия

E-mail: fedotov@sbras.ru

МОДЕЛЬ ИНФОРМАЦИОННОЙ СИСТЕМЫ ДЛЯ ПОДДЕРЖКИ НАУЧНО-ПЕДАГОГИЧЕСКОЙ ДЕЯТЕЛЬНОСТИ *

Описывается технологический подход к созданию модели информационной системы для поддержки научно-педагогической деятельности, организованной в виде электронной библиотеки, а также архитектура информационной системы и принципы интеграции с цифровым депозитарием, правила представления и преобразования метаданных. Основное внимание уделяется работе со словарями ключевых терминов, которые используются для систематизации и классификации информационных ресурсов, и моделированию связей с фактами.

Ключевые слова: информационная система, электронная библиотека, словарь-справочник, классификация информационных ресурсов, база данных, цифровой депозитарий, информационно-поисковый тезаурус, ключевые термины, DSpace, протокол OAI-PMH, метаданные.

Введение

Развитие технологий в области передачи и обработки информации, в частности создание современных телекоммуникационных систем, привело к появлению принципиально новых возможностей организации практически всех этапов научно-образовательного процесса, что, в свою очередь, обусловило качественный рост информационных потребностей ученых и преподавателей. Современный подход к организации работ с документами и материалами связан в первую очередь с мировой тенденцией перевода разнородной информации с бумажных носителей в цифровую форму и с созданием крупномасштабных информационных хранилищ. Представление информации и знаний в электронной (цифровой) форме позволяет принципиально по-иному создавать, хранить, организовывать доступ и использовать информацию. Наряду с этим формируется новый класс информационных систем, предназначенных для управления электронными информационными ресурсами, – электронные библиотеки [1; 2]. При этом возникает целый ряд новых задач, в их числе – организация разграниченного пользо-

* Работа выполнена при частичной поддержке РФФИ (проекты №12-07-00472, 13-07-00258), президентской программы «Ведущие научные школы РФ» (грант НШ.–5006.2014.9).

Федотов А. М., Барахнин В. Б., Жижимов О. Л., Федотова О. А. Модель информационной системы для поддержки научно-педагогической деятельности // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2014. Т. 12, вып. 1. С. 89–101.

ISSN 1818-7900. Вестник НГУ. Серия: Информационные технологии. 2014. Том 12, выпуск 1

вательского доступа к собранной информации и предоставления пользователю непосредственно самой информации.

Современные информационные технологии предоставляют исследователю мощный аппарат для «манипулирования данными», а не информацией. Данные, переведенные в электронную форму, приобретают новое качество, обеспечивая им более широкое распространение и эффективное использование. На первый взгляд, может сложиться впечатление, что развитие информационных технологий уже само по себе способно вывести работу с научной информацией на качественно новый уровень, но, к сожалению, это совсем не так, поскольку пока нет адекватного аппарата для оперирования с «информацией» и информационными ресурсами [3; 4]. Сами по себе данные (как набор битов) не представляют никакой информационной ценности без соответствующих описаний или моделей. Применение информационных технологий должно основываться на использовании различных моделей (феноменологических, информационных, математических и др.). Как неоднократно отмечал А. А. Ляпунов (см., например, [5]): «нет модели – нет информации». Для возможности продуктивной работы нужны данные, превращенные в «информацию», представленную в виде «знаний» – «адекватного отражения действительности в сознании человека в виде представлений, понятии, суждений теорий».

В процессе научной, а особенно педагогической, деятельности много времени и сил отнимает работа с литературными источниками и документами (информационными ресурсами особого типа): поиск необходимых документов, систематизация и классификация в соответствии с поставленной задачей. В настоящее время существуют достаточно мощные информационные системы, которые в той или иной степени удовлетворяют информационные потребности пользователей [6]. Однако основными недостатками большинства систем являются ограниченность возможностей проведения аналитической работы с ресурсами и обеспечения интеграции ресурсов как внутри каждой из систем, так и с внешними системами (низкая интероперабельность) [4]. Это крайне неудобно в сфере научно-образовательной деятельности, одна из главных задач состоит в том, что необходимо установить связи между конкретными научными фактами (например, «что означает термин кибернетика» или «кто автор данной статьи») и сущностями информационной системы (персоны, факты, документы, публикации и т. п.).

Далее будем использовать следующее понимание факта: *«входящая в текст документа характеристика сущности, описываемой в онтологии информационной системы, представляемая как единичное значение данных».* Факт может быть извлечен из информационного содержания объекта либо определен экспертом. Факт может определять как свойства (атрибуты) объекта, так и его связь с другими объектами.

В монографии [7], изданной ВИНИТИ еще в 1976 г. и содержащей подробный обзор теоретических проблем информационного поиска, на основе выделения двух типов информационных потребностей – потребности в сведениях об источниках необходимой научной информации и потребности в само́й необходимой научной информации – говорится, что для удовлетворения информационных потребностей первого предназначены информационные системы, получившие название документальных, второго типа – фактографических. В настоящее время наиболее востребованным средством информационного обеспечения научной деятельности становятся интеллектуальные информационные системы (ИИС), сочетающие возможности информационных систем обоих названных типов и позволяющие удовлетворять информационные потребности квалифицированного пользователя в соответствии со схемой «документ – факт – рассуждение» [4; 8]. В дальнейшем мы будем использовать понятие «фактографические системы» в широком смысле, включающем и ИИС.

В интеллектуальных информационных системах в качестве составного компонента выступают рассуждающая система, формализующая правила логического вывода, и интеллектуальный интерфейс. Документы, с которыми приходится работать в процессе научно-образовательной деятельности, 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 и факты из электронных документов достаточно произвольной структуры. Ее решение позволит получать новую информацию и знания [4].

Уточнение понятия «факт»

Подробный анализ значения термина «факт» и его производных, основанный на соответствующих статьях «Философской энциклопедии» и «Словаря современного русского литературного языка», был проведен в монографии [7]. В итоге были выявлены следующие признаки фактов.

1. Факты следует отличать от данных, фиксирующих специфику объекта, условия наблюдения и т. п. Понятие же научного факта «предполагает элиминирование такой информации, т. е. требует определенного обобщения непосредственных данных». Однако при этом отмечается, что четкого различия между указанными понятиями в «Словаре современного русского литературного языка» не приводится.

2. Фактом можно назвать лишь знание, выдержавшее критическую проверку, т. е. полученное в результате обобщения и переработки данных абстрактно-логическим мышлением (разумеется, при этом надо отдавать отчет в том, что достижение абсолютно достоверного знания является лишь идеалом развития науки, практически недостижимым).

3. Любой факт, прежде чем стать объектом научной коммуникации, должен быть преобразован в текст или изображение и получить форму научного документа или его части. Более того, «объектом сбора, хранения, поиска и выдачи в так называемых фактографических информационно-поисковых системах… могут быть лишь соответствующие тексты или документы, описывающие некоторые данные или факты, если под документом понимать… любой фрагмент такого текста» [7].

Нетрудно видеть, что сформулированные признаки весьма расплывчаты. Прежде всего, признаки 1 и 2 предполагают обобщение и оценку перерабатываемых данных. Поэтому жесткое соблюдение требований, вытекающих из указанных признаков, выводит работу с фактами за рамки собственно научно-образовательной деятельности, поскольку в той или иной степени требует использования теорий и методик конкретных научных дисциплин, к которым относятся данные.

Трудно провести четкую границу между фактами и непосредственными данными. Это касается следующих типов сущностей, описывающих тот или иной объект исследования: имена собственные, хронологические сведения, различные характеристики объектов и т. п.

В монографии [9] отмечается, что *данные* понимаются в ней как факты и идеи, представленные в символьной форме, позволяющей проводить их передачу, обработку и интерпретацию, а *информация* – как смысл, приписываемый данным на основании известных правил представления фактов и идей. Структурированная (связанная причинно-следственными и иными отношениями) информация, образующая систему, составляет *знания*.

Для уточнения смысла, вкладываемого в термин «факт» применительно к информационным системам, представляется целесообразным использование семиотического подхода, подобно тому, как это было сделано в работе [8] для терминов «информация», «знание», «тезаурус», «онтология». В этой работе, в частности, показано, что данные соответствуют синтаксическому уровню сообщения (в том числе документа), информация (в узком смысле!) – семантическому, а знания – прагматическому.

Отметим, что на такое понимание понятия «факт» обратил внимание Людвиг Витгенштейн. Понятие «факт» является центральным в его «Логико-философском трактате» [10], одним из источников которого, как отметил Л. Витгенштейн предисловии трактата, стали работы Г.Фреге – основателя семиотики. Процитируем основные положения трактата, касающиеся фактов:

«…1.1. Мир есть совокупность фактов, а не вещей.

...

1.2. Мир распадается на факты.

1.21. Любой факт может иметь место или не иметь места, а все остальное останется тем же самым.

....

2. То, что имеет место, что является фактом, – это существование атомарных фактов.

2.01. Атомарный факт есть соединение объектов (вещей, предметов).

2.011. Для предмета существенно то, что он может быть составной частью атомарного факта.

…

2.034. Структура факта состоит из структур атомарных фактов.

2.04. Совокупность всех существующих атомарных фактов есть мир.

2.05. Совокупность всех существующих атомарных фактов определяет также, какие атомарные факты не существуют.

2.06. Существование или несуществование атомарных фактов есть действительность. (Существование атомарных фактов мы также называем положительным фактом, несуществование – отрицательным.)

2.061. Атомарные факты независимы друг от друга.

2.062. Из существования или несуществования какого-либо одного атомарного факта нельзя заключать о существовании или несуществовании другого атомарного факта.

…

4.21. Простейшее предложение, элементарное предложение, утверждает существование атомарного факта.

…

4.22. Элементарное предложение состоит из имен. Оно есть связь, сцепление имен».

Положения, выдвинутые в «Логико-философском трактате», имеют большое значение для семиотики, в частности, потому, что в нем устанавливается полное соответствие между онтологическими и семантическими понятиями [11]. Кроме того, Витгенштейн не исключает ложные (или, если угодно, представляющиеся на данном уровне познания ложными) утверждения из числа атомарных фактов, а называет такие факты несуществующими.

Нетрудно заметить, что процитированные положения «Логико-философского трактата» (прежде всего, ключевые определения из раздела 2.01: «Атомарный факт есть соединение объектов (вещей, предметов)... Структура факта состоит из структур атомарных фактов») практически полностью воспроизводятся в модели данных Чена «сущность – связь» [12], являющейся основой для унификации различных представлений данных.

Для единообразия определения понятия «факт» удобно использовать модификацию модели данных «сущность – связь» из той же статьи, называемую моделью множества сущностей. Ее отличительные особенности заключаются в том, что, во-первых, в ней всё трактуется как объекты (в том числе, например, цвет; в то время как в модели «сущность – связь» цвет обычно трактуется как «значение», а согласно «Логико-философскому трактату» «2.0251. Пространство, время и цвет (цветность) есть формы объектов»), а во-вторых, все связи в этой модели – бинарные. Связи между объектами в модели множества сущностей также рассматриваются как объекты, связанные, в свою очередь, с объектами – атрибутами связей.

Важно подчеркнуть, что создание фактографических систем подразумевает извлечение фактов не только непосредственно из текста документа, но и из его метаданных. Это следует, например, из традиционного понимания научно-информационного процесса [13], 2-й этап которого (аналитико-синтетическая переработка документальной информации) предусматривает как извлечение сведений о содержании документа (индексирование, аннотирование и т. п.), так и обработку его библиографических данных.

Заметим, что указание источника, из которого извлечен данный факт, в качестве одного из атрибутов факта позволяет с той или иной степенью достоверности отделять «существующие» (в терминологии Витгенштейна) факты от «несуществующих». С этой целью на множестве источников может быть введена шкала их достоверности.

Таким образом, отношения и факты объявляются первичными, а вещи представляют собой пересечение, совокупность возможных отношений. Другими словами, с вещью можно соотнести общую область «пересечения» множества фактов. Атомарный факт есть соединение (двух) объектов. Анализ фактов дает объекты или предметы. При этом по мере накопления фактов представление о вещи может меняться. Благодаря такой трактовке мира вещь выступает не как нечто данное, застывшее, вполне определенное, а как некоторая сущность с размытыми границами, и эти границы уточняются по мере выявления класса возможных для данной сущности отношений (фактов). Чтобы определить вещь, надо зафиксировать все факты (положительные, где может встречаться эта вещь, и отрицательные, где не может).

Отсюда вытекает, что функционирование интеллектуальной информационной системы основано на двух противоположных процессах: при пополнении ИС новыми сведениями происходит преобразование семантической информации в данные, однако непосредственно потребности пользователя удовлетворяет обратный процесс – извлечение из данных нужной пользователю информации и знаний.

Ввиду того, что информация в ИС отображает некоторые сущности (предметы, процессы, явления, персоны, публикации, факты, ключевые термины и т. п.), следует рассматривать информационную систему как множество информационных объектов – наборов данных, представляющих (описывающих) эти сущности в ИС.

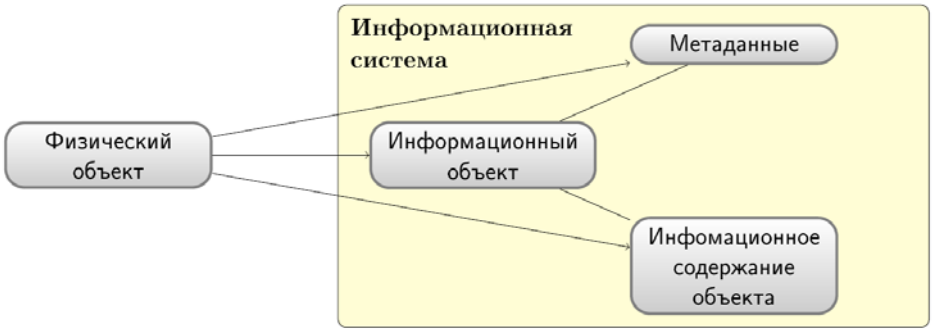

Рис. 1. Структура информационной системы

Эффективным средством описания информационных объектов в ИС являются метаданные – данные, являющиеся неотъемлемой частью информационного объекта и описывающие реальный объект или группу объектов (рис. 1).

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» [4; 14].

Под термином *электронная библиотека* (ЭБ) в данной работе будем понимать систему управления структурированными каталогизированными коллекциями разнородных электронных (цифровых) документов. ЭБ предоставляет средства навигации и поиска (в отличие от печатных изданий, микрофильмов и других носителей). ЭБ способна не только обеспечить многосторонний поиск и навигацию в каталогах, 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.

Электронные библиотеки – явление относительно новое, но уже достаточно популярное [15]. Тем не менее ЭБ сегодня следует рассматривать как множество слабосвязанных сущностей, объединяемых, на первый взгляд, только общим названием [2; 17]. 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 [17]. 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 [18]:

- управление информационными ресурсами;
- 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 обеспечиваются только посредством визуальных графических интерфейсов. Это хорошо для пользо-

вателя-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 [1].

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 [18]. 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, [19]) показывает их разнородность на уровнях:

- информационной модели, которую они обеспечивают;
- поддержки пользователей и групп пользователей;
- 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 [1]. Наиболее важным является профиль метаданных информации, циркулирующей в системе. Выбор профиля должен основываться на выполнении следующих требований:

- включать 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 возможность 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,

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

электронный объект с реализованными связями уже не совсем соответствует своему печатному оригиналу. Во-вторых, внедренные в объект связи должны быть гарантированно 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.

Дополнительно отметим, что в информационные объекты ЭБ могут содержать информацию, которая не имеет к объектам хранения традиционных библиотек никакого отношения [16]. Речь может идти:

- об электронных копиях элементов хранения традиционных архивов;
- об изображениях элементов хранения традиционных музеев;
- о видео-, аудиоинформации, полученной разными способами, например, видеозапись доклада, сделанного на конференции;
- о научных или других фактах;
- и т. д. и т. п.

Для решения возникающих проблем необходимо использовать концептуальные модели, обобщающие накопленный опыт в сфере создания и использования ЭБ [19].

Существует достаточно много технологических разработок информационных систем для электронных библиотек, так или иначе ориентированных на поддержку научных исследований. Среди них следует отметить информационные системы, близкие к фактографическим, например ИСИР (ЕНИП) РАН [20], ИРИС СО РАН [21], euroCRIS [2], и документальные, например eLibrary [3], Информика [4], MathNET [5]. Названные системы в той или иной степени удовлетворяют потребности исследователей в информации, однако каждая из них имеет определенные недостатки.

Функциональные требования к модели электронной библиотеки

Как уже отмечалось, 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.

В информационной системе, направленной на поддержку научно-образовательной деятельности, важно хранить описание жизненного цикла информационных ресурсов и иметь возможность восстановить состояние ресурса на любой момент времени. Существуют информационные ресурсы, которые должны 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 их восстановления.

Как отмечалось, другой важной проблемой является идентификация информационных ресурсов [18; 22], 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 до очень подробных описаний информационных объектов.

[2] http://www.eurocris.org/
[3] http://elibrary.ru/
[4] http://www.informika.ru/
[5] http://www.mathnet.ru/

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 слабоструктурированных 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 [23; 24]: хранилище данных – репозиторий, сервер метаданных, сервер приложений, словари-справочники (см. рис. 2).

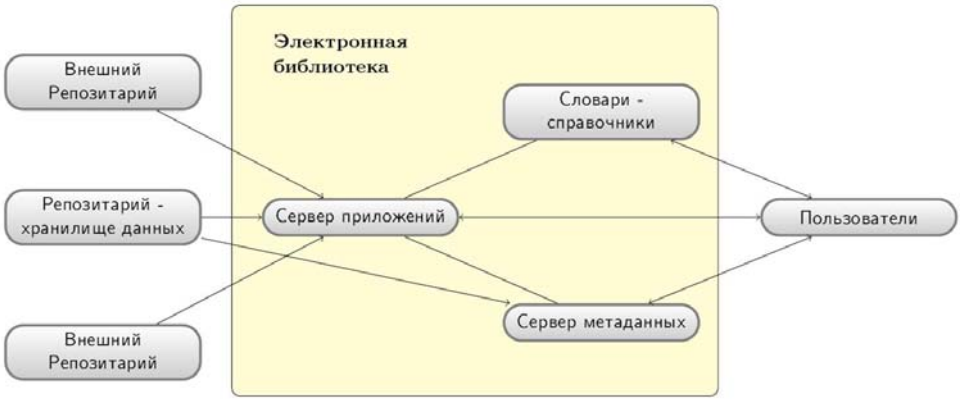

Рис. 2. Архитектура электронной библиотеки

Выбор метаданных для ЭБ

Для поддержки сложных функций поиска и классификации необходимы: поддержка поиска по атрибутам, полнотекстового поиска, а также просмотр ресурсов по категориям и словарям-классификаторам.

В существующих ИС информационные ресурсы разрозненны, не очень хорошо 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. Под интероперабельностью ИС 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. Это приводит к требованию соответствия всех схем данных, интерфейсов и протоколов международным стандартам и рекомендациям [1; 18].

В работах [18; 24] был определен профиль ЭБ как необходимый набор стандартов и компонентов информационной системы, ориентированной на научные исследования.

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 системы метаданных и форматы метаописаний, а также разные подходы к решению прикладных задач. Решением этих проблем занимаются многие организации во всем мире. Большое внимание им уделяют такие международные организации, как W3C [6], DCMI [7], OCLC [8], IFLA [9], IETF [10], ISO [11].

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.

Как мы уже отмечали, основу содержания ЭБ составляют информационные объекты, которые представляют следующие основные типы сущностей:

- субъекты (персоны, организации и т. п.);
- объекты (единицы хранения – публикация, документ, факт, научный результат, мероприятие, фотография и др.);
- отношения (понятие, ключевой термин, событие, время, место).

В отличие от общепринятых документных (библиографических) ЭБ указание на субъекты дается ссылкой на экземпляр сущности субъект, что позволяет корректно решать задачу идентификации объектов.

Используемый профиль определяет список элементов данных (полей), необходимых для создания записи соответствующего типа, и раскрывает содержание элементов данных. 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.

Для формирования метаданных применяются несколько стандартов, являющихся расширениями рекомендаций Dublin Core [12] и Qualified Dublin Core (QDC). Для документов нами была расширена стандартная схема метаданных QDC полями, включающими основные требования государственного стандарта МЕКОФ [13].

Словари (ключевые признаки, ключевые термины) – 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. Необходимо рассматривать различные типы ключевых терминов, а именно:

- ключевые термины в стандартном понимании;
- ключевые термины, описывающие персону;

---

[6] World Wide Web Consortium (W3C) – http://www.w3.org/

[7] DCMI – Dublin Core Metadata Initiative (http://www.dublincore.org/).

[8] Online Computer Library Center – http://oclc.org/

[9] International Federation of Library Associations – http://www.ifla.org/

[10] Internet Engineering Task Force (IETF) – http://www.ietf.org/

[11] International Organization for Standardization – http://www.iso.org/

[12] http://www.dublincore.org/

[13] ГОСТ 7.19-2001. Система стандартов по информации, библиотечному и издательскому делу. Формат для обмена данными. Содержание записи.

- ключевые термины, описывающие организацию;
- ключевые термины, описывающие временные периоды;
- ключевые термины, описывающие географические понятия, а также тематические словари-классификаторы, тезаурусы, описания предметной области данной научной школы и классификаторы документов в соответствии с МЕКОФ.

Имеется ряд российских (например, УДК [14], ГРНТИ [15]) и международных (например, MSC-2000 [16], ORTELIUS [17]) словарей для классификации научных данных. Однако в целом эти словари содержат только общенаучную информацию и не годятся (использовать их все равно надо) для систематизации материалов.

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.

Проблемы классификации

Наиболее распространенным вариантом классификации документов является фасетная классификация, теория построения которой формализована индийским библиотековедом Ш. Р. Ранганатаном [25]. Объекты классифицируются одновременно по нескольким независимым друг от друга признакам (фасетам). Применительно к ЭБ (и электронным ресурсам вообще) в качестве фасетов выступают элементы метаданных, в том числе и ключевые термины.

Важно отметить, что при создании научно-образовательных ЭБ, для которых библиографические признаки документов гораздо менее важны по сравнению с обычными ЭБ, подмножества множеств значений библиографических метаданных, образующих значения фасетов, как правило, более широки. Так, ссылки на различные переиздания одного и того же документа с точки зрения научно-образовательных ЭБ целесообразно считать эквивалентными.

Простейшая формальная модель классификации документов с использованием структурированных метаданных документов выглядит следующим образом [14]. В каталоге ЭБ хранится информация о документах d_i. При этом любой документ d_i каталога системы представляется как $d_i = < m_i^{j,k} >$, где $m_i^{j,k}$ – значения элементов метаданных M_j, k – количество значений (с учетом повторений) соответствующего элемента метаданных в описании документа. Рассмотрим подмножество метаданных M_C, определяющее набор классификационных признаков документов, используемых для составления поискового предписания (с учетом заданных логических операций). Для фиксированного элемента метаданных M_j из подмножества M_C множество документов разбивается на классы эквивалентности, соответствующие различным значениям или же заранее выбранным подмножествам множества значений этого элемента метаданных.

Будем считать два документа *толерантными*, если у них совпадает значение хотя бы одного из элементов метаданных, входящих в M_C (напомним, что толерантность – отношение, которое обладает свойствами рефлексивности и симметричности, но, вообще говоря, может не обладать, в отличие от отношения эквивалентности, свойством транзитивности). Каждое такое значение порождает класс толерантности [26].

Рассмотрим всевозможные сочетания значений элементов метаданных, входящих в M_C. Множества документов, обладающие одинаковым набором значений, суть ядра толерантности, которые служат классами эквивалентности на множестве документов. С содержательной точки зрения этой ситуации соответствует вхождение некоторого раздела классификатора ЭБ в раздел более высокого уровня, когда оба этих раздела учитываются при описании про-

[14] УДК – Универсальная десятичная классификация.
[15] ГРНТИ – Государственный рубрикатор научно-технической информации.
[16] MSC-2000 – Математический классификатор – http://www.ams.org/msc/msc.html
[17] The «Ortelius Thesaurus on Higher Education» – http://cordis.europa.eu/cerif/src/sum_concl.htm

странства толерантности (разумеется, можно и не учитывать раздел более низкого уровня при определении толерантных элементов, но тогда мы будем иметь дело с пространством толерантности, отличным от первоначального).

Таким образом, поисковое предписание, содержащее подмножества метаданных, определяющее набор классификационных признаков и сочетаний значений этих метаданных при помощи логических операций, определяет конкретное ядро толерантности на множестве документов, которое и выдается пользователю в качестве ответа на его информационный запрос. На множестве классов толерантности также можно, в свою очередь, ввести отношение толерантности, при этом толерантными считаются классы, имеющие хотя бы один общий документ. Такая конструкция оказывается полезной, например, для организации поиска документов «по аналогии».

Формализм, основанный на использовании отношения толерантности, оказывается более удобным при создании ЭБ, поскольку в отличие от обычных библиотек, в которых классификаторы заданы априорно, при работе с ЭБ нередко приходится применять те или иные алгоритмы кластеризации документов (см., например, [4]), а уже потом, исходя из результатов кластеризации, устанавливать подмножества множеств значений элементов метаданных, выступающих в качестве значений фасетов.

Важно отметить, что при создании научно-образовательных ЭБ, для которых библиографические признаки документов гораздо менее важны по сравнению с обычными ЭБ, подмножества множеств значений библиографических метаданных, образующих значения фасетов, как правило, более широки. Так, ссылки на различные переиздания одного и того же документа с точки зрения научно-образовательных ЭБ целесообразно считать эквивалентными.

Практическая реализация

Рассмотренная модель информационной системы представлена в виде электронной библиотеки, реализующей учебные пособия [18] по курсам «Современные проблемы информатики и вычислительной техники», «Вычислительные системы», «Информатика» и «Экология» и др.

Основной каталог информационных ресурсов сервера метаданных информационной системы строится в соответствии со схемой метаданных МЕКОФ. Для долговременного хранения документов используется репозиторий DSpace [19]. Стандартная схема метаданных DSpace, основанная на схеме DCMI, расширена полями, отвечающими основным требованиям МЕКОФ. Для поддержки процесса наполнения полнотекстовых баз созданные профили метаданных зарегистрированы в системе DSpace и в соответствии с ними настроены рабочие процессы, а также пользовательский интерфейс системы.

Для организации обмена метаданными между DSpace и сервером метаданных (а также с другими системами с расширенным профилем) создан специальный сервис, выполняющий преобразование метаданных из внутренней схемы DSpace в другие схемы метаданных, в том числе и в схему DCMI с использованием квалификаторов (QDC [20]), а также в схему МЕКОФ (представление ISO2709 или XML). Реализован также OAI-PMH сервис [21], который в пакетном режиме периодически, в соответствии с расписанием, проводит синхронизацию метаданных репозитория и сервера метаданных. Для заполнения основного каталога метаданных в соответствии с созданными схемами метаданных используются контролируемые словари из справочного блока сопровождения. Для обеспечения интероперабельности данных DSpace также задействован сервер приложений на основе ZooPARK-ZS [27], обеспечивающий доступ к метаданным системы по протоколам Z39.50 [28] [22] и SRW/SRU [23].

[18] http://fedotov.nsu.ru/lecture.php

[19] http:// www.dspace.org.

[20] Qualified Dublin Core (QDC) – http://www.dublincore.org/documents/dcmi-terms/

[21] The Open Archives Initiative Protocol for Metadata Harvesting [Электронный ресурс]: Protocol Version 2.0 of 2002-06-14 // Open Archives Initiative: [web-сайт] / The OAI Executive; OAI Techical Commitee. 2004 (http://www.openarchives.org/).

[22] ANSI/NISO Z39.50-2003. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. NISO Press, Bethesda, Maryland, U.S.A. November 2002.

[23] SRU (Search/Retrieve via URL) URL: http://www.loc.gov/standards/sru/ (дата обращения: 23.08.2013).

Разработанная модель информационной системы может быть использована как типовая модель системы для работы с документами, связанными с научно-образовательной деятельностью, поскольку решает основные задачи, предъявляемые к этим системам: обеспечение системы надежного долговременного хранения цифровых (электронных) документов с сохранением всех смысловых и функциональных характеристик исходных документов; обеспечение «прозрачного» поиска и доступа пользователей к документам, как для ознакомления, так и для анализа содержащихся в них фактов; организация сбора информации по удаленным цифровым репозиториям, поддерживающим протоколы OAI-PMH, SRW/SRU, Z39.50.

Список литературы

1. *Жижимов О. Л., Мазов Н. А., Федотов А. М.* Некоторые заметки об эволюции цифровых репозитариев традиционных библиотек к полнофункциональным электронным библиотекам // Вестн. Владивосток. гос. ун-та экономики и сервиса. Территория новых возможностей. 2010. Т. 7, № 3. С. 55–63.

2. *Антопольский А. Б., Вигурский К. В.* Концепция электронных библиотек // Электронные библиотеки: рос. науч. электронный журн. 1999. Т. 2, вып. 2. URL: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/1999/part2/antopol (дата обращения: 04.05.2013).

3. *Федотов А. М.* Парадоксы информационных технологий // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2008. Т. 6, вып. 2. С. 3–14.

4. *Шокин Ю. И., Федотов А. М., Барахнин В. Б.* Проблемы поиска информации. Новосибирск: Наука, 2010.

5. *Ляпунов А. А.* О соотношении понятий материя, энергия и информация // Ляпунов А. А. Проблемы теоретической и прикладной кибернетики. Новосибирск: Наука, 1980. С. 320–323.

6. *Барахнин В. Б., Федотов А. М.* Исследование информационных потребностей научного сообщества для построения информационной модели описания его деятельности // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2008. Т. 6, вып. 3. С. 48–59.

7. *Михайлов А. И., Черный А. И., Гиляревский Р. С.* Научные коммуникации и информатика. М.: Наука, 1976.

8. *Барахнин В. Б., Федотов А. М.* Уточнение терминологии, используемой при описании интеллектуальных информационных систем, на основе семиотического подхода // Изв. вузов. Проблемы полиграфии и издательского дела. 2008. № 6. С. 73–81.

9. *Арский Ю. М., Гиляревский Р. С., Туров И. С., Черный А. И.* Инфосфера: Информационные структуры, системы и процессы в науке и обществе. М.: ВИНИТИ, 1996.

10. *Витгенштейн Л.* Логико-философский трактат. М.: Изд. иностр. лит., 1958.

11. *Грязнов А. Ф.* Витгенштейн. Новая философская энциклопедия. М.: Мысль, 2000. Т. 1. С. 406–408.

12. *Чен П. П.-Ш.* Модель «сущность – связь» – шаг к единому представлению данных // СУБД. 1995. № 3. С. 137–158.

13. *Михайлов А. И., Черный А. И., Гиляревский Р. С.* Основы информатики. М.: Наука, 1968.

14. *Федотов А. М., Барахнин В. Б.* Проблемы поиска информации: история и технологии. Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2009. Т. 7, вып. 2. С. 3–17.

15. *Земсков А. И., Шрайберг Я. Л.* Электронные библиотеки: Учеб. пособие. 3-е изд. М.: ГПНТБ России, 2004.

16. *Воройский Ф. С.* Электронные и традиционные библиотеки – суть не одно и то же // Электронные библиотеки: рос. науч. электронный журн. 2003. Т. 6, вып. 5. URL: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part5/voroisky (дата обращения: 04.05.2010).

17. *Акимов С. И., Елизаров А. М., Ершова Т. В., Когаловский М. Р., Федоров А. О., Хохлов Ю. Е.* Научно-методическая поддержка разработки научных электронных библиотек // Электронные библиотеки: рос. науч. электронный журн. 2005. Т. 8, № 1. URL: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2005/part1/AEEKFH.

18. *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.

19. *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.

20. *Бездушный А. Н., Бездушный А. А., Серебряков В. А., Филиппов В. И.* Интеграция метаданных Единого научного информационного пространства РАН. М.: Вычислительный центр им. А. А. Дородницына РАН, 2006.

21. *Шокин Ю. И., Федотов А. М. Жижимов О. Л., Гуськов А. Е., Столяров С. В.* Электронные библиотеки – путь интеграции информационных ресурсов Сибирского отделения РАН // Вестн. Казан. нац. ун-та. 2005. № 2, спец. выпуск. С. 115–127.

22. *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.

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

24. *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, вып. 2. С. 5–14.

25. *Ранганатан Ш. Р.* Классификация двоеточием. Основная классификация / Пер. с англ. М.: ГПНТБ СССР, 1970.

26. *Шрейдер Ю. А.* Равенство, сходство, порядок. М.: Наука, 1971.

27. *Жижимов О. Л., Федотов А. М., Шокин Ю. И.* Технологическая платформа массовой интеграции гетерогенных данных // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2013. Т. 11, вып. 1. С. 24–41.

28. *Жижимов О. Л., Мазов Н. А.* Принципы построения распределенных информационных систем на основе протокола Z39.50 / ОИГГМ СО РАН. Новосибирск: ИВТ СО РАН. 2004. 361 с.

Дата поступления в редколлегию 17.03.2014

A. M. Fedotov, V. B. Barakhnin, O. L. Zhizhimov, O. A. Fedotova

A MODEL OF INFORMATION SYSTEM TO SUPPORT SCIENTIFIC AND EDUCATIONAL ACTIVITIES

The paper describes a technological approach for developing a model of information system to support the scientific and educational activities, organized in the form of a digital library. The information system architecture, the principles of integration with digital depository and the rules of metadata representation and transformation are described. Special emphasis is put on work with dictionary of key terms, that are used for information resources organization and classification, and on modeling relations with the facts.

Keywords: information system, digital library, glossary, information resource classification, database, digital depository, information retrieval thesaurus, key terms, DSpace, protocol OAI-PMH, metadata.