Построение модели информационной системы для описания научных школ СО РАН*

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

Аннотация

В статье предлагается подход к созданию биографо-библиографической модели информационной системы для описания научных школ СО РАН. Важнейшим требованием к к такой информационной системе является идентификация документов и субъектов. Удовлетворение этого требования достигается тем, что формировании метаданных того или иного ресурса используются авторитетные базы данных, с помощью которых устанавливаются конкретные ссылки на субъекты (персоны). Вторым важным требованием является организация поиска с привлечением онтологии. Для его реализации требуется дополнительная информация о предметной области, включающая определения терминов, сущностей и связей. Представление этой информации должно соответствовать глобальным договоренностям международным стандартам, иначе поиск с использованием онтологии будет ограничен текущей системой, а интероперабельность не будет реализована.

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

1 Научные школы СО РАН

Научные школы — ценное интеллектуальное наследство СО РАН.

Изучение научного наследия основателей сибирских научных школ является важной задачей. Период их деятельности совпал со временем стремительного развития науки. Сегодня отчетливо вырисовывается истинное значение идей и событий, масштабность личностей. Среди сибирских ученых есть всемирно признанные, сделавшие весомый вклад в мировую науку. Главное достоинство СО РАН как раз в том, что оно сохранило научные школы. Многие лидеры ушли из жизни, а научные школы остались и продолжают свое развитие. Школы живут, потому что остались ученики, не разрушены традиции и сохранилась память. Нашему поколению оставлено богатейшее научное наследие, которое

*Работа выполнена при частичной поддержке РФФИ: проекты 08-07-00229, 09-07-00277, 10-07-00302, президентской программы «Ведущие научные школы РФ» (грант № НШ 6068.2010.9) и интеграционных проектов СО РАН.

следует сберечь. Промедление в этой работе может привести к невосполнимым утратам, связанным с временным фактором: потерей документов, уходом из жизни очевидцев событий.

Одним из первоочередных способов решения задачи сохранения научного наследия является комплекс мероприятий, направленный на быструю и качественную публикацию материалов наследия в виде электронных библиотек в сети Интернет.

2 Электронные библиотеки

Электронная библиотека — структурированная каталогизированная коллекция разнородных электронных документов (в отличие от печатных изданий, микрофильмов и других носителей), снабженная средствами навигации и поиска. Электронные библиотеки — явление достаточно новое и достаточно популярное, но тем не менее, электронные библиотеки сегодня следует рассматривать как множество слабосвязанных сущностей, объединяемых на первый взгляд только общим названием.

Под термином «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 [1].

Однако, задача электронной библиотеки — не только обеспечить многосторонний поиск в каталоге, но и предоставить пользователю непосредственно найденный ресурс (публикацию, фотографию, описание научного факта и др.), а также дополнительные сведения о нем, например, об авторах редакторах, библиографии, организации и т. п. Важным фактором электронных библиотек является определение метаданных для описания ресурсов и выделение ключевых видов субъектов и объектов.

В существующих информационно-поисковых системах, когда сведения о ресурсах представлены в виде слабоструктурированного текста и полнотекстовый поиск нужных данных осуществляется по запросам в свободной форме, пользователь получает огромное количество «шумовой» информации, среди которой очень трудно выбрать действительно полезные знания. Учитывая это обстоятельство, для представления сведений о ресурсах необходимо использовать каталогизацию ресурсов, структурное представление и метаданные, описывающие содержимое ресурса в виде набора именованных значений, в том числе связей с другими ресурсами [2, 3].

Метаданные используются для автоматизированного анализа содержимого ресурса, построения поисковых индексов и позволяют обеспечить достаточно высокую точность и эффективность поиска разнородной информации. Эти требования приводят к необходимости создания специализированных информационных систем, обличенных в форму электронных библиотек (ЭБ) [4, 5], позволяющих решить основные проблемы интеграции разнородных распределенных информационных ресурсов на основе технологий и принципов построения открытых систем [6].

Отметим, что основу разработки электронной библиотеки составляют, прежде всего,

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 являются профили 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 открытыми, т. е. обеспечивать доступ к соответствующей информации по этим описаниям;
- быть расширяемыми, 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 ориентированными на современные и перспективные технологии описания и использования информации;
- обеспечивать возможности интероперабельности с внешней средой.

Для формирования простых метаданных применяются несколько стандартов, являющимися расширениям рекомендаций Dublin Core [8].

С точки зрения потребностей научных сотрудников существенным недостатком многих схем метаданных электронных библиотек является то, что они работают лишь с так называемыми документоподобными объектами, определяют метаданные, описывающие только такие ресурсы, не выделяют другие виды важных объектов, например, персоналии, организации, коллекции и т. п. В итоге, например, встретив упоминание персоны в одном месте, невозможно точно установить соответствие с ее упоминанием в другом месте. Это обусловлено тем, что метаданные рассматриваются как нечто, связанное только с документом, их используют как средства идентификации ресурсов только для документов и только для целей их извлечения. Решение этой проблемы может быть достигнуто следующим способом. При формировании метаданных того или иного ресурса (при его каталогизации) необходимо использовать авторитетные базы данных (авторитетные файлы), с помощью которых устанавливать конкретные ссылки на объекты [9].

Отметим, что при работе с цифровыми объектами человечество уже выработало определенный набор стереотипов, отсутствие которых вызывает дискомфорт [10]. Одним из

элементов этого набора является требования 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. Это уже другой объект. Во-вторых, внедренные в объект связи должны быть гарантировано актуальными. Никого, например, не интересуют гиперссылки, ссылающиеся на несуществующие документы. Так появляется требование обеспечения ссылочной целостности данных. Это очень жесткое требование, которое тяжело 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 электронной библиотеке 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;
- об изображениях элементов хранения традиционных музеев;
- о видео-, аудио- информации, полученной разными способами, например, видеозапись доклада, сделанного на конференции;
- о научных или других фактах;
- и т.д. и т.п.

3 Идентификация документов и субъектов

В существующих реализациях ЭБ, как правило, не решается проблема идентификации документов и субъектов, поскольку метаданные рассматриваются только для целей описания документов.

Решение этой проблемы может быть достигнуто следующим способом. При формировании метаданных того или иного ресурса (в процессе его каталогизации) необходимо использовать авторитетные базы данных (авторитетные файлы), с помощью которых устанавливать конкретные ссылки на персоны.

Применение технологии авторитетного контроля записей само по себе решает проблему идентификации персон. Однако иногда связи между авторитетными и библиографическими записями отсутствуют или становятся некорректными (в условиях объединения нескольких каталогов, в каждом из которых используются свои коды авторитетных записей) и далеко не всегда описываются все персоны, имеющее отношение к ресурсу. Другой проблемой, связанной с идентификацией персоны, является в целом неудовлетворительное качество авторитетных записей, связанных с обеспечением научной деятельности.

В традиционной библиотеке возможности поиска ограничивались поиском по алфавитному или систематическому каталогам информационных ресурсов с прямой ссылкой (указания шифра хранения) на сам ресурс. Использование электронных каталогов расши-

рило поисковые возможности, но сохранило основным типом поиска поиск по предопределенным поисковым атрибутам. Это атрибутивный поиск, именно этот тип поиска сегодня является основным в традиционных библиотеках, в том числе и в библиотеках цифровых объектов. Фактически при этом поиск производится только по массивам вторичной информации, оставляя открытым вопрос соответствия последней первичным информационным ресурсам.

Другой возможный тип поиска — поиск по заданным шаблонам. Наконец, поиск с привлечением онтологии является поиском более интеллектуальным, для его реализации требуется дополнительная информации — информация о предметной области, включающая определения терминов, сущностей и связей. Следует отметить, что представление этой дополнительной информации должно соответствовать глобальным договоренностям — международным стандартом, иначе, поиск с привлечением онтологии всегда будет ограничен текущей системой, а интероперабельность не будет реализована [3].

Использование публикаций (информационных ресурсов) в научно-исследовательском процессе выдвигает необходимость быстрого ознакомления с содержимым публикации, и аннотации здесь может оказаться недостаточно. В связи с этим должны быть разработаны средства полуавтоматического выделения оглавления и выделения фактов (научных результатов в соответствии с онтологией, понятиями) с обеспечением ссылок на соответствующие разделы документа, а также средства работы с библиографическими ссылками. Наиболее важным для описания научной школы являются отношения, связанные с предметной областью понятиями (концептами).

Правильный авторитетный контроль информационных ресурсов должен давать конкретное указание на персоны с учетом их отношений к данному ресурсу: автор, редактор, персонаж, владелец, рецензент и т. д., что позволит корректно решать задачу идентификации объектов.

4 Типы сущностей ЭБ

Основу содержания ЭБ составляют следующие основные типы сущностей:

- субъекты: люди (персоны) и организации;
- объекты — единицы хранения: произведение, выражение, воплощение, экземпляр, факт, научный результат, мероприятие, фотография и др.;
- отношения: понятие, событие, время и место.

Субъекты и Объекты-хранения описываются схемами данных в соответствии с профилем ЭБ. Каждая единица хранения снабжается максимально подробным описанием в соответствии с ГОСТом, включая коды известных классификаторов (например, ББК, УДК, Current Contents, MSC2000), а также ссылками на понятие из словарей-классификаторов ЭБ.

При возможности дается ссылка на репозиторий (полнотекстовую БД). В отличие от общепринятых библиографических БД указание на субъекты: персоны (автор, редактор, персонаж и т.д.) и организации (институт, университет, издательство, журнал и т. д.)

дается ссылкой на экземпляр сущности субъект, что позволяет корректно решать задачу идентификации объектов.

Отметим, что помимо общепринятых описательных метаданных основные сущности электронной библиотеки должны быть снабжены именованными отношениями, из которых можно выделить следующие:

- Входит в состав (ссылка: объект) — данный ресурс является физически или логически частью указанного ресурса.
- Включает (ссылка: объект) — данный ресурс физически или логически включает указанный ресурс.
- Работал (ссылка: субъект; атрибут: время)
- Преподавал (ссылка: субъект; атрибут: время)
- Изображен (ссылка: объект; атрибут: время)
- Ученик (ссылка: субъект)
- Автор (ссылка: объект; атрибут: время)
- Персонаж (ссылка: объект; атрибут: время)

В качестве примеров таких отношений, представленных в ЭБ, можно привести, например:

- документ, описывающий книгу и совокупность документов, описывающих отдельные главы этой книги;
- описание конференции и список докладов или список презентаций на этой конференции.

Такого сорта отношения между документами моделируется путем задания связей тип «родитель-потомок». Последний вид связи, реализованный в модели ЭБ, определяет списки подчиненных документов в зависимости от условий истинности заданных администратором предикатов.

Для связи с предметной областью используется тезаурус. Понятия (словарь-онтология), которые имеют отношение к конкретной научной школе, строятся как подмножество тезауруса предметной области, дополненное словарями из предметных указателей монографий. Важным дополнением стандартного тезауруса являются списки (словари) фактов: научных достижений научных школ с соответствующими отношениями. Планируемая система является персоноцентричной: все объекты и отношения, понятия, факты, мероприятия, публикации и др. привязываются к персонам.

Стандартная схема данных персоны расширяется биографическим описанием (или ссылками на биографические описания) и свойствами связывающие данную персону с научной школы и основными понятиями из тезауруса научной школы.

Основные свойства (атрибуты) структуры отношения персона-объект следующие:

- принадлежит (школе) (атрибуты: время, место);
- участвует (мероприятие, событие и т. п.) (атрибуты: время, место);
- имеет (награды, звания и т. п.) (атрибуты: время, место);

- создатель (публикация, результат, школа и т. п.) (атрибуты: время, место);
- действующее лицо (персонаж) (атрибуты: время, место).

Основные свойства (атрибуты) структуры отношения персона-персона следующие:

- учитель;
- ученик.

Принадлежность персоны к научной школе определяется либо отношением учитель-ученик («генеалогическим древом»), либо отношением совместной работы с основателем научной школы или с другими членами научной школы, определяемым местом работы и научными результатами.

Построение «генеалогического древа» предполагается вести двумя способами:

1) автоматически по базе данных авторефератов (учеником считается тот, у кого данная персона была руководителем кандидатской или консультантом по докторской диссертациями – ученики ищутся не только у основателя школы, но и у его учеников);

2) вручную, используя информацию «очевидцев».

Для создания полнофункциональной информационной системы необходимо расширить набор метаданных для описания публикаций (в первую очередь, авторефератов и диссертаций) такими значениями атрибутов:

- оппоненты;
- научная новизна;
- пристатейный список публикаций.

5 Заключение

В работе представлен подход к созданию модели информационных систем по научным школам. Главной задачей при проектировании ЭБ является построение хорошей модели, учитывающей мировой опыт подобных разработок. Важным фактором при создании ЭБ является определение метаданных для описания ресурсов и выделение ключевых видов субъектов и объектов. Одной из важнейших задач идентификации документов является задача создания авторитетных файлов. При этом следует решить задачу реализации алгоритма автоматического анализа авторитетных файлов, позволяющего связывать библиографические и авторитетные записи. С целью быстрого ознакомления с содержимым публикации в научно-исследовательском процессе должны быть разработаны средства полуавтоматического выделения оглавления и выделения фактов (научных результатов в соответствии с онтологией, понятиями) с обеспечением ссылок на соответствующие разделы документа, а также средства работы с библиографическими ссылками. Построение тезауруса для связи с предметной областью понятиями.

Список литературы

[1] *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.

[2] *Федотов А. М., Барахнин В. Б.* Проблемы поиска информации: история и технологии // Вестник НГУ. Серия: Информационные технологии. — 2009. — Т. 7. — Вып. 2. — С. 3–17.

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

[4] *Федотов А.М.* Концептуальные подходы к построению распределенных систем // Труды Международной конференции по вычислительной математике МКВМ-2004. — Новосибирск: Изд. ИВМ и МГ СО РАН, 2004. — С. 132–143.

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

[6] ISO/IEC 7498-1:1994. Information technology - Open Systems Interconnection - Basic Reference Model: The Basic Model.

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

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

[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 // Вестник НГУ. Серия: Информационные технологии. — 2011. — Т. 9. — Вып. 1. — С. 89–101.

[10] *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. — №3 (7). — 2010. — С. 55–63.