Электронные архивы, музеи и экспозиции

Марчук А.Г. (Новосибирск)

Институт систем информатики им. А.П.Ершова СО РАН
mag@iis.nsk.su

Аннотация

В работе рассматриваются особенности построения информационных систем частного вида: электронные архивы и музеи. Показано, что системы данного класса в информационном плане являются открытыми системами. Предложен подход к созданию таких систем, отличающийся тем, что используется три концепта: информационная модель предметной области, информационная модель мира и информационная тема. Анализируются особенности реализации электронных архивов и музеев.

1 Введение

Архивы и музеи обладают двойной функциональной нагрузкой: первое – сохранять «единицы хранения» и предоставлять их для работы и, второе – всей совокупностью экспонатов и конкретными экспозициями раскрывать одну или множество познавательских тем. Для архивов первая функция, как правило, является более значимой, для музеев – более значимой является вторая.

Не вдаваясь в технические детали «виртуализации» архивных и музейных объектов, электронный архив или музей, в первую очередь, являются коллекцией достаточно однородных документов, являющихся информационными образами предметов. Отношение к этому виду информационных систем как к замкнутому и однородному информационному образованию – достаточно типично и присутствует в работах многих авторов [1].

Однако легко показать, что раскрытие темы или тем, как основная функция для большинства таких коллекций, не реализуемо только документами коллекции. Действительно, за «единицами хранения» стоят люди, события, факты, интерпретации и т.д. В ряде коллекций, например книжных библиотеках, отражение сущностей реального мира сознательно вынесено за рамки информационных (библиографических) систем, кроме ограниченного их числа: авторы, издательства и т.д. Это обосновано тем, что книга, как источник знаний, как правило, структурой своего построения и внутренним содержанием раскрывает какую-то тему или группу тем, а внешняя информация дается через ссылки. Для архивных и музейных объектов – это не так. Для них всегда имеется необходимость раскрытия некоторых тем, причем на одних и тех же экспонатах могут раскрываться темы из самых разных областей. Например, на фондах Эрмитажа, могут представляться и раскрываться такие темы как: творчество конкретного художника, художественное направление, шедевры, их стоимость, книги об искусстве, аукционы, история похищений и возвращений и т.д. и т.п. Разными будут: набор экспонатов, их структура, наборы вспомогательных материалов.

Для более удобной работы с архивными или музейными материалами требуется создавать открытые информационные системы. Открытые в том смысле, что база данных единиц хранения должна сочетаться с более общей базой данных, отражающей сущности реального мира. В идеале, база данных общих сущностей должна быть внешней по отношению к электронному хранилищу. Должна осуществляться лишь привязка данных коллекции к общемировому полю фактографических данных. Однако сейчас, приходится воспроизводить кусочек модели мира в рамках конкретных информационных систем. Поскольку единого подхода к таким «сущностным» базам данных не выработано, разные «кусочки» информационной модели мира не объединяются в единое информационное пространство. Все мы заполняем многочисленные регистрационные формы на себя лично, на свою организацию, не имея возможности указать идентификаторы «своих» данных или вовсе их не указывать.

Созданием подходов к построению информационной модели мира занимается множество исследователей. Автор данного доклада предлагает свой подход, заключающийся в том, что сущности реального мира разбиваются на (непересекающиеся) категории, которых небольшое количество: система, свойство, изменение, сценарий и роль. Информационные элементы разных категорий могут находиться в некоторых отношениях, например, какую-то роль может играть конкретная система. Элементы одной категории могут принадлежать некоторым классам, классы

Труды 5ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2003, Санкт-Петербург, Россия, 2003.

упорядочены по иерархии наследования свойств (атрибутов).

При таком подходе простые знания о мире (факты) относительно легко отобразить в информационную модель мира, а ее использовать для «привязки» к ней документов, составляющих электронную коллекцию. Более, того, коллекция, как частный элемент, легко описывается метасхемой модели мира.

При наличии такой структуры информационной модели, возможно описание темы раскрытия для порождения виртуальной экспозиции. Такой запрос на экспозицию сродни поисковому запросу, типичному для многих информационных систем. Разница лишь в том, что необходимо также задать и принцип структурирования результирующей информации. Гипотетически, возможны генераторы экспозиций, выдающие в автоматическом режиме как саму (виртуальную) экспозицию, так и ее оформление.

2 Электронные архивы и музеи

Стартуя от «вещной» коллекции будь то архив документов и предметов или набор экспонатов, артефактов и т.д., достаточно просто перейти к электронному варианту коллекции через сканирование, компьютерный набор, ведение каталога, ввод учетных карточек. Процесс этот, хотя и трудоемок, но в первом приближении – понятен и рутинен. Специфические трудности порождения электронного образа экспоната преодолеваются специалистами и не являются предметом данной работы. Нас будет интересовать проблема оформления всей, в общем случае большой, коллекции в информационную систему, помогающую пользователям пользоваться, а информационным администраторам – пополнять и модифицировать данные.

Таким образом, первичной основой электронного музея или архива является все то же множество «единиц хранения» теперь уже оформленного как набор документов (ресурсов). Очевидна и мультимедийность такого набора – это тексты, гипертексты, базы данных, фотографии, аудио и видеозаписи, списки, схемы, планы, 2D и 3D-модели и т.д. Для работы такой набор должен быть оформлен в виде, дающем возможности как изучения каждого документа (через группу файлов, отражающих этот документ), так и поиска нужного документа в соответствии с некоторыми критериями. Вторая задача становится все более актуальной при росте числа экспонатов. Если с коллекцией, составляющей десятки единиц, еще можно работать через общий список, реальные архивы содержат десятки-сотни тысяч и миллионы документов/экспонатов.

При разработке концепции электронного архива академика А.П.Ершова были учтены следующие факторы:

- Сам Андрей Петрович, будучи аккуратным и системным человеком, сохранил максимально возможное количество существенных документов и расположил документы в некоторой системе, отражающей виды деятельности, события и временные этапы.
- Деятельность А.П. отражена весьма разнообразна как по направлениям (научная, педагогическая, административная, общественная, литературная и др.) так и по форме (переписка, служебные записки, приказы, черновики работ, организационные документы и т.д.)
- Ершов по роду деятельности и по свойствам характера имел контакты с очень большим числом людей, представляющих разные организации, города и страны. Общение велось на нескольких языках (естественно, главные – русский и английский)
- Информация ершовского архива интересна людям разных профессий и возрастов, живущих в разным местах по земному шару.

Анализ этих факторов и сопоставление альтернатив привели разработчиков к группе структурных и технологических решений, наиболее существенными для настоящей работы являются:

- Основной структуризацией документов выбрано группирование А.П.Ершова.
- База данных информационной системы состоит из основной базы данных (картотеки) документов, дополнительных баз данных персон и организаций, а также ряда вспомогательных таблиц: города, страны, должности, языки и др.
- Дополнительная структуризация формируется через привязку документов к дополнительным базам данных.
- Обращено особое внимание на удобство интерфейсов пользователей (Front end) и информационных администраторов (Back end).

В настоящее время электронный архив академика А.П.Ершова технологически полностью создан, работа по его наполнению все еще продолжается (пока переведено в электронное представление только около 50% документов архива). Архив выставлен в Интернет и эксплуатируется около двух лет. Опыт эксплуатации в основном подтвердил правильность архитектуры и заложенных идей. В частности, замечательным свойством созданной информационной системы является возможность «серфинга» по документам архива как по тематической группировке, так и через упомянутых людей и организации. Другой интересной реализованной возможностью является подсистема

генерации справки о персонах по материалам документов и баз данных архива.

Настало время переработать систему с учетом опыта, новых идей и появившихся стандартов.

3 Анализ возможностей и некоторых современных подходов

Как уже указывалось, первичным информационным полем являются карточки на документы/экспонаты. Их можно, и видимо нужно, оформлять в соответствии с рекомендациями Dublin Core (DC) [2] и другими международными стандартами. Это порождает первичную возможность включения архива/музея в мировое информационное пространство.

Поля, определяемые DC являются полями двух видов – свойства и отношения. Поля-свойства описывают единицу хранения и характеризуют содержание (Title, Creator, Subject, Description, Publisher и др. Небольшое количество полей-отношений (Identifier, Type, Source, Relation) устанавливают связи между единицами хранения и «внешним» миром.

В практическом плане рекомендациям DC не хватает формализации для того, чтобы можно было единообразно смотреть на информационные ресурсы разных коллекций. Причинами этого является традиционная «очеловеченность» взгляда на библиографические и аналогичные системы и, второе, сложность задачи структуризации разнородной информации, описывающей реальный мир.

Новую методологию построения информационных систем несут в себе концепции и стандарты группы Semantic Web [3]. Это в первую очередь относится к стандартам RDF и RDFS и появляющейся технологической формализации онтологического подхода (OWL [4] и др.).

Базовая идея предложения в том, чтобы метаинформацией описывать не только свойства описываемой сущности, но и элементы смысла сущности через установление регламентированных связей или отношений между информационными объектами (ресурсами). Другими ключевыми элементами подхода являются: первое – формат данных ориентирован изначально не на человека, а на машинных (программных) агентов. Вторым элементом является понятие концептуализации, формализующей модель предметной области, относительно которой создается ИС.

В итоге, информационный ресурс, напр. коллекция, публикуется в информационном пространстве как пара: концептуализация-данные, выполненная в подходящем формализме, напр. RDFS-RDF. Агент, использующий такую базу данных может работать в той же модели предметной области или другой («своей»). В последнем случае имеется возможность создать медиатор, согласующий разные представления. Недостаток такого подхода проявляется когда множество (n) агентов пытаются для своих целей использовать множество (m) разнородных баз данных. Необходимость иметь в этом случае n*m медиаторов существенно снижает ценность подхода. Выходом, естественно, является порождение для всех поставщиков семантических данных общей модели, которую естественно назвать моделью мира.

Легко показать трудность создания модели мира, удовлетворяющей различным требованиям и условиям. Разумно подойти к этой задаче построив огрубленную модель мира, формулирующую инвариант по категориям сущностей и понятий относительно предметных областей. Тогда концептуализация разных баз данных будет выглядеть как модель мира + модель предметной области, выраженной в согласованных с моделью мира понятиях.

4 Отражение сущностей реального мира

Информационные системы в целом и отдельные информационные элементы в частности, призваны отражать реальный мир. Неструктурирующее отражение (типа фотографии) не несет в себе даже зачатков смысла отражаемого материала, его смысл формируется в «голове» у интерпретатора (человека, животного, программы) в соответствии с моделью интерпретации и контекстом отражения. Нас будут интересовать структурирующие отражения, выносящие часть модели интерпретации (а значит – смысла), в отражаемую информацию.

Простейшим способом структуризации окружающего мира является выделение целостных образований, позволяющих про «это» говорить как о едином. Соответственно вещи, предметы, объекты и т.д., попадают в первичную структуризацию именно по этому критерию, а экземпляры однородных множеств одинаковых целостностей, различаются через именование. Такой вид структуризации существенно расширяет свои возможности через добавление в него отношения общее-частное. Собственно такой подход и дал название «классификация» и закреплен во многих методологиях, включая объектно-ориентированное программирование (концепция классов и наследования).

Через иерархию классов целостность получает цепочку имен типа: объект – вещь – деталь – гайка – гайка ГОСТ 45051-65 М13 – «вот эта самая гайка с зазубриной». Классификационные схемы отрабатываются и используются в самых разных предметных областях и охватывают как материальные предметы, так и множество абстракций, таких как виды деятельностей, разделы наук и знаний, способы и алгоритмы и т.д. и т.п.

При том, что классификационный подход успешно применяется и развивается, его одного не хватает для порождения информационного пространства, обладающего свойством поисковой удобности. Это связано с многоярусностью,

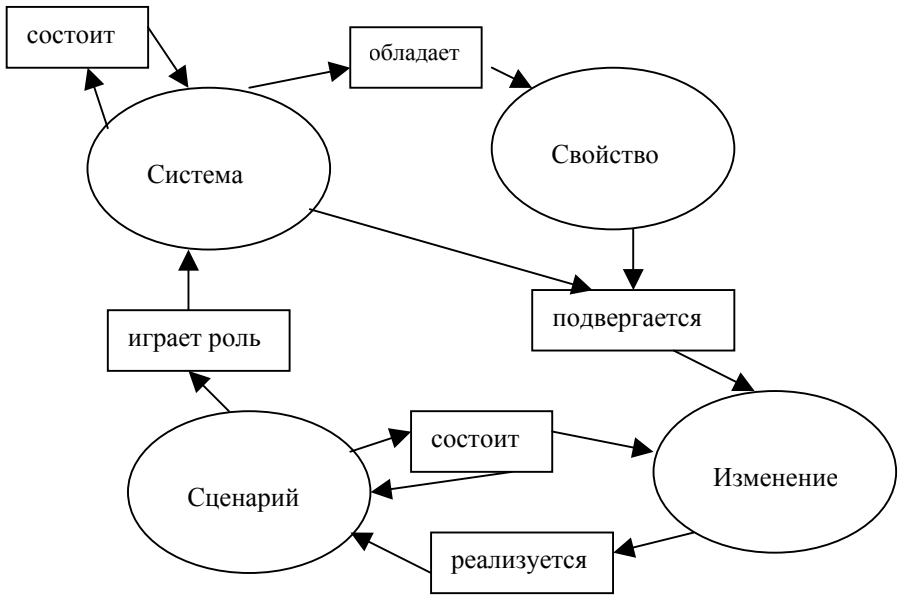


Рисунок 1

неочевидностью и неоднозначностью классификаторов. Кроме того, «многогранные» объекты и явления, такие как человек, сложный объект, событие, процесс и др., не укладываются в единственный классификатор, а должны характеризоваться множеством разных аспектов.

Предлагается построить модель мира через выделение небольшого количества категорий сущностей с достаточно жесткой дисциплиной отношений между категориями.

На рисунке 1 изображена метамодель предлагаемой модели мира. Овалами обозначены категории сущностей: система, свойство, изменение и сценарий. Прямоугольниками – отношения между сущностями.

Под системой понимается нечто «вещное» и целостное, напр. предмет, человек, документ. Изменение предполагает абстракцию «предыдущее состояние – последующее состояние». Изменение всегда осуществляется посредством сценария как группы изменений и подсценариев (иногда называемых ролями). Роли в сценариях играют системы. Например, в изменении «образование молодых людей» эффективен сценарий «университет», роль которого может осуществлять конкретная организация (напр. НГУ). В свою очередь, в сценарии «университет» предполагается роль (множество действий-обязанностей и подсценариев) «ректор», роль которого исполняет конкретный Имярек.

Такая модель все еще довольно «груба» для целей описания семантики сущностей, но она может быть развита до любой степени детальности через два дополнительных механизма. Первым расширением является деление указанных категорий сущностей на непересекающиеся или слабо пересекающиеся классы исторически сложившихся видов.

Возможным разделением может быть следующее:

Система
- предмет
- существо
- человек
- географическая система
- документ
- конструкция
- организация

Изменение
- изготовление
- модификация
- разрушение
- отражение
- измерение
- вычисление, перевод

Сценарий
- событие
- деятельность
- организация
- семья
- проживание
- награды, степени, звания, дипломы
- гражданство
- имущественные отношения

Второе расширение осуществляется за счет использования отношения частное-общее и соответствующей системы классов систем и сценариев. При детализации такой модели мира мы выходим на неоднозначности, отражающие неоднозначности и неопределенности наших представлений о мире, а также разнообразие вариантов традиций, правил и условностей. Предполагается, что можно зафиксировать более малую часть расширения как (огрубленную) модель

мира, оставив дальнейшие уточнения на модель предметной области. По крайней мере, в этом случае, модель предметной области будет согласована с моделью мира.

5 Выборки, поиск, информационная тема, экспозиция

При предложенном подходе общая база информационной системы состоит из информационных единиц разных видов и отношений между ними. При наличии большого количества таких единиц, отражающих разнородную информацию, выделение нужной части или отдельного элемента является опорным действием для большинства оформительских или редактирующих преобразований.

Семантические сети, используемые в Semantic Web и конкретных технологиях, например RDF, родственны подходу, основанному на реляционной алгебре, соответственно, типовые выделяющие запросы оформляются в подходящем формализме, например XQuery. В любом случае, это фильтр, устанавливающий отношения среди элементов локального контекста, определяемого искомым элементом. Локальный контекст определяют центральный элемент контекста и информационные элементы, непосредственно прилегающие к нему по имеющимся в базе данных отношениям. Например, если центральным элементом поиска является документ, то в локальный контекст также могут входить: информация о документе – автор, дата создания, вид документа и т.д., отражение или упоминание каких элементов базы данных имеется в тексте документа, какую роль играет (играл) документ в разных сценариях (событиях, организациях), в каких других документах имеется упоминание о данном и т.п.

Собственно, такое формирование поискового образа сродни человеческому, поскольку память человека в большой мере базируется на ассоциациях, т.е. отношениях между разными сущностями.

Легко видеть, что установление контекста может быть интерактивной задачей поскольку к центральному элементу может примыкать очень большое количество элементов его потенциального контекста и сужение такого пространства можно производить через вовлечение в контекст более «удаленных» элементов. Например, представим себе, что в документе может упоминаться персона «Иванов Иван Иванович». Но в силу распространенности Ивановых и в силу того, что данные о нем могут быть сокращенными (Иванов И.И. или просто Иванов) или на другом языке, для указания корректного отношения в контексте, надо выделить из общей базы данных конкретного Иванова, задав подходящий дополнительный контекст, например: Иванов, проживавший в городе Денвер, штат Колорадо и участвовавший в конгрессе IFIP 1988 года.

Выделяющие, в частности, вложенные выделяющие запросы хорошо отработаны для реляционного подхода как в логическом, так и техническом аспектах. Вообще, семантические сети того класса, который рационален для отображения модели мира по всей видимости соответствуют реляционной системе таблиц и могут ими реализовываться. Проблема в другом – интерфейс конечного пользователя для формулирования контекста. Конечному пользователю не рационально представлять поисковый инструмент в виде формализованного алгебраического языка запросов. В этом плане, наиболее интересным представляется использование вариантов подхода Query-by-example.

Как уже указывалось, для (виртуальный) музеев принципиальным вопросом является отображение информации о единицах хранения и прилегающей к ним контекстной информации в экспозицию, отражающую некоторую информационную тему. С учетом уже сказанного, видно, что информационная тема также задается некоторым поисковым запросом определяющим степень общности предметной области (напр. «Дифференциальные уравнения в частных производных»), временные и географические рамки, виды выдаваемых информационных элементов, существенные для данной темы поля и др. Существенные в формулировании поискового фильтра могут быть особенности аудитории (язык, возраст, образование, цель и т.д.).

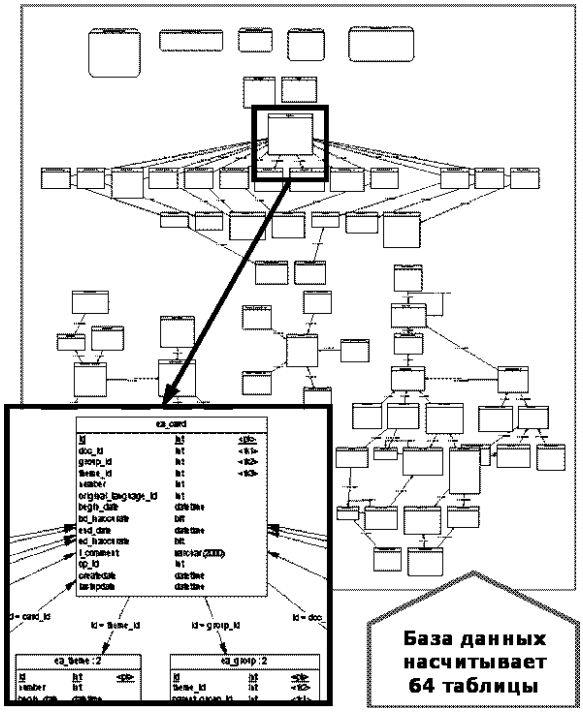
Экспозиция это не только набор информационных единиц, но и структура информации о каждой, структура экспозиции, навигация и поиск в экспозиции. Для многих случаев, базовой структуризацией экспозиции может быть древовидное построение, соответствующее традиционному «книжному». Такое выстраивание материала легко сделать имея шаблоны выдачи информации разного вида (люди, организации, события, документы и т.д.) и производя рекурсивную развертку (гипер)текста по шаблонам относительно прямых отношений между элементами выборки. Естественно, элементы, присутствующие в экспозиционной выборке через разные отношения или рекурсивные отношения или элементы специальных видов (напр. статьи и книги), не будут включаться текстуально в содержание более «главной» статьи, а будут оформляться в отдельные «страницы», доступные через гиперссылки.

Понятно, что в области оформления экспозиции могут быть также реализованы самые разнообразные подходы от формализованной справки о предмете запроса до виртуальных «залов», «гидов» и т.д.

6 Особенности организации электронного архива А.П.Ершова

Предложенный подход был сформирован как результат работы над электронным архивом академика А.П.Ершова и опыта эксплуатации этого электронного архива [5, 6]. В созданном электронном архиве, кроме образов бумажных страниц (элементов хранения) имеется три важных базы данных: темы, люди и организации. Привязка электронных карточек, описывающих документы к элементам имеющихся баз данных позволяет качественно по-иному работать с документами архива через «серфинг» по ссылкам и группам выборок. Удалось также частично реализовать раскрытие простых тем типа «ученый имярек в документах архива Ершова».

В качестве характеризации важности привязки



документов архива к внешним сущностям, приведем короткую статистику по сформированной базе данных (обработаны 251 папок из 520):

Число введенных документов	18167
Число графических образов страниц документов	60267
Количество персон в базе данных	4344
Количество городов в базе данных	680
Количество стран	49
Количество языков, использованных в документах	15

Видно, что при таком информационном разнообразии, использование сопряженных баз данных вносит новые поисковые и структурирующие возможности.

Как уже указывалось, ядром электронного архива А.П.Ершова является база данных. Детали базы данных не являются существенными для рассмотрения данной работы, а общая схема реляционных таблиц данных приведена на рисунке 2.

Заключение

В настоящее время предлагаемый подход используется в следующем этапе развития электронного архива А.П.Ершова и в работе по археологическому и этнографическому порталу.

Работа поддержана грантом РФФИ и частично финансируется интеграционным проектом Сибирского отделения РАН.

Литература

[1] Когаловский М.Р. Энциклопедия технологий баз данных. – М.: Финансы и статистика, 2002

[2] Dublin Core Metadata Element Set Reference Description, Version 1.1, 1999-07-02. http://purl.org/dc/documents/rec-dces-19990702.htm

[3] Rossi G., Schwabe D., Lyardet F. Web Application Models are more than Conceptual Models. // In Proc. Workshop on the WWW and Conceptual Modeling (WWWCM'99), pp. 239-253, Springer 2001

[4] Patel-Schneider P.F., Hayes P., Horrocks I. OWL Web Ontology Language, Semantics and Abstract Syntax. // W3C Working Draft. http://www.w3.org/TR/owl-absyn/

[5] .Marchuk A., Nemov A., Fedorov R., Antyoufeev S. The Information System for Creating and Maintaining the Electronic Archive of Documents // Lect. Notes Comput. Sci. – 2002. – Vol. 2457. – P. 175-184

[6] Антюфеев С.В и др. Проектирование, создание и наполнение электронного архива // Электронные библиотеки: перспективные методы и технологии, электронные коллекции / Труды четвертой Всерос. конф. Дубна, 2002. – Дубна: ОИЯИ, 2002. – Т. 2. – С. 189-196.

Digital Archives, Museums and Expositions

Alexander Marchuk

The paper is concerning to different aspects of building information systems at archive and museum resources. It is shown, that systems of this class are open systems in information sense. New approach to design of such systems is proposed. Systems should be based on three concepts: semantic data model of digital collection, semantic data model of universe and information theme.

Also particular features of digital archives and museums are analyzed. Experience in design and support of A.P.Ershov digital archive is discussed.