Подходы к работе с хранилищами слабоформализованных данных

© Якшин М. М.

БЕН РАН
greycat.na.kor@gmail.com

Аннотация

Данный документ посвящен проблеме хранения и представления слабоструктурированных данных. Рассматриваются преимущества и недостатки традиционных подходов и возможность применения Wiki-технологий для организации подобных хранилищ.

1 Введение

В последнее время всё большую актуальность приобретают задачи создания всевозможных масштабных хранилищ данных, и одно из наиболее сложных и неоднозначных направлений в этом вопросе – хранение слабоформализованных данных с произвольной, меняющейся в зависимости от задач, структурой.

Когда идёт работа со строго формализованным данными, обычно используются стандартные решения на реляционных базах данных, но зачастую возникает задача по обработке некоего произвольного множества данных, как-то связанных между собой. Некоторые из этих данных имеют форму произвольных полнотекстовых документов, некоторые – документов с четко выраженной структурой (разделами, подразделами), а некоторые – наборы структурированных данных, с разбиением по полям вплоть до атомарных значений.

Одним из важных шагов к решению этой задачи стало повсеместное внедрение стандартов, базирующихся на XML, а также технологий RDF и Semantic Web.

2 Традиционные подходы

Сама по себе задача сейчас не имеет четко обоснованного и теоретически красивого решения. Все универсальные подходы обладают теми или иными достоинствами и недостатками. Исторически, один из самых первых и самых простых способов создания связанной системы документов – это массив гипертекстовых файлов (с появлением стандарта HTML/XHTML – это файлы в этом формате) с расставленными ссылками друг на друга. Создание хранилища данных на этой основе возможно, но у этого решения существует несколько серьезных проблем:

1) Высокий порог вхождения. Для обычного пользователя создание даже одного HTML-документа – серьезная задача, которая требует неких специальных знаний и, обычно, привлечение WYSIWYG-средств (дополнительного программного обеспечения) разной степени сложности. Создание же системы связанных между собой документов вручную, даже для компетентного специалиста, становится уже сложной задачей, для решения которой, как правило, привлекаются специальные средства, называемые Content Management System – но они, в свою очередь, урезают универсальность подхода до решения какой-то конкретной задачи – например, ведения веб-сайта с новостными лентами, или ведения просто архива статей по датам.

2) Смешение оформления и содержания. Изначально HTML – язык произвольной разметки произвольных данных, включал в себя как семантически-нагруженные элементы (например, strong или em), так и элементы только для оформления, такие, как font (шрифт), b (полужирный), i (курсив), u (подчеркивание) и т.п. Особенно при использовании средств WYSIWYG, пользователю в, первую очередь, предоставляются именно такие тэги.

Таким образом, для создания системы связанных документов слабой формализации одних только механизмов, предоставляемых XML и XHTML, недостаточно. Как правило, используются некие внешние средства автоматизации деятельности по созданию такой информационной коллекции, некие CMS, которые следят за всем массивом документов и хранят их в качестве какого-то единого репозитария.

3 Технологии Wiki

В последнее время получила распространения технология Wiki – технология, разработанная в конце 90-х годов на первой волне популяризации Internet. Изначально wiki-системы представляли собой просто сайты, страницы которых мог редактировать кто угодно, прямо из web. В современном варианте – это система для сбора и структурирования информации, характеризующаяся следующими признаками:

1) Многопользовательский режим работы – все редактирование осуществляется через web-интерфейс, есть центральный сервер (или кластер), хранящий весь массив данных.

2) Возможность многократно править текст посредством самой wiki-среды (вебсайта), без применения особых приспособлений на стороне пользователя.

3) Проявление изменений сразу после их внесения.

4) Разделение информации на однозначно идентифицируемые документы.

5) Несложный человеко-читаемый язык разметки, позволяющий легко отделить содержимое от оформления.

6) Учёт изменений (учёт версий) текста и возможность отката к ранней версии.

Преимущества Wiki:

1) Именование и идентификация. Wiki представляет собой коллекцию произвольных документов, единственный способ доступа к которым – идентификатор. В самом документе ссылка на другой документ создается автоматически.

2) Шаблоны – предоставляют возможность хранения и представления структурированных данных, подробнее будут рассмотрены ниже.

3) Автоматическое создание всевозможных списков и классификаций. Один из примеров реализации таких механизмов – это механизм категорий или тэгов, применяющийся часто в разных wiki. Он заключается в том, что документ помечается, как принадлежащий какой-то категории. После этого при обращении к документу "Категория:Заданное имя" (в пространстве имен "Категория") будет выведен список документов, помеченных заданным именем.

4) Ссылочная целостность. Как правило, wiki предоставляет возможность отследить как все ссылки с текущей страницы куда-либо, так и все ссылка откуда-то на текущую страницу.

Во многих wiki (в том числе в mediawiki, которая рассматривается в качестве основного инструмента в настоящей статье), существует механизм создания шаблонов. Шаблон – это отдельный документ, физически по способу хранения ничем не отличающийся от статей, но находящийся в отдельном пространстве имен и доступный для вставки с помощью специальных тэгов. Как правило, вызов шаблона из основного документа имеет вид вызова некоей процедуры, которой передается набор параметров – его и можно рассматривать как набор сильноструктурированных данных внутри произвольного документа.

Шаблон играет двоякую роль: с одной стороны – его вызов хранится в теле документа и таким образом формирует некую структуру машинно-обрабатываемых данных, которые в дальнейшем могут индексироваться, их можно анализировать, преобразовать при необходимости в отдельную реляционную форму, строить по ним статистики и т.п.

Существует и другой, альтернативный подход к созданию такого хранилища слабоструктурированных данных, который предполагает использование двух раздельных XML, привязанных к документу для хранения слабоструктурированной информации. Первый такой фрагмент XML – это собственно сами данные, а второй – стандартное XSLT-преобразование, которое будет применяться к этому фрагменту для показа его конечному пользователю. Этот способ – более гибкий, но он сложнее, так как требует создания трех отдельных документов.

Для модельной реализации механизма хранения слабоструктурированных сведений на основе wiki для проекта электронной библиотеки «Наследия России» автором была создана серия шаблонов, который в качестве аргументов получает некое подмножество наиболее распространённых полей библиографического описания (например, монографического издания – Шаблон:Книга), а на выходе – для пользователя – выдает простейшее библиографическое описание по ГОСТ 7.1-2003. При этом же, для возможного последующего машинного анализа, по сути сохраняется всё разбиение данных по полям: не теряется информация.

Стоит заметить, что почти все аргументы могут быть как просто текстовыми элементами, так и ссылками.

В качестве недостатков предложенного метода можно отметить чрезмерную простоту создаваемых отношений: создаваемые отношения хотя и классифицируются как "много-ко-многим", но не имеют возможности задания каких-либо атрибутов у самой связи. Частично это реализуется с помощью создания связей из разных полей шаблона, что позволяет задать тип отношения между сущностями (например, "автор", "издательство" и т.п.), но не позволяет задать, например, временных рамок существования отношения или каких-либо более сложных атрибутов. Однако, стоит заметить, что эта простота точно так же является и огромным преимуществом для конечного пользователя – когда схема данных будет описана программистом, задача ввода и редактирования данных значительно упрощается.

Таким образом, предложенная схема хранения данных может использована в системах электронных библиотек, где требуется организовать работу со слабоструктурированными данными, в частности, в системе, предоставляющий доступ к развернутому описанию авторов и предметов исследования электронной библиотеки «Наследия России».

Литература

1] Ward Cunningham, Bo Leuf. The Wiki Way. Quick Collaboration on the Web. – Boston, Addison-Wesley: 2001. – ISBN 0-201-71499-X.

2] Anja Ebersbach, Markus Glaser, Richard Heigl, Gunter Dueck. Wiki: Web collaboration. – Berlin, New York: Springer, 2006. – ISBN 3-540-25995-3.

Approaches to organizing semistructured data repositories

Yakshin, M.M.

This document describes the problem of storage and presentation of semistructured data. This work addresses advantages and disadvantages of traditional approaches and compares them to Wiki-technologies used to organize such repositories.