Ермаков Денис Евгеньевич
Институт программных систем РАН, Исследовательский центр медицинской информатики, г. Переславль-Залесский

В настоящей статье рассматриваются современные тенденции в области представления и передачи информации в медицинских информационных системах. Определяется, что одним из наиболее перспективных направлений с точки зрения повышения эффективности передачи и обработки является структурирование медицинской информации. Даётся обзор наиболее распространённых форматов представления структурных данных, а также анализ опыта их применения в медицинской информатике.

Введение

Вне зависимости от содержания медицинского документа, свобода врача в выражении своего мнения делает ситуацию со стилистикой документации достаточно пёстрой даже в пределах одного лечебного учреждения, не говоря уже об отличиях, существующих в этом вопросе между разными ЛПУ. В самом общем случае врач может быть полностью свободным в выборе формы своего отчёта, а также может использовать практически любые выбранные по своему усмотрению термины.

Однако существует множество доводов в пользу структурирования документации [1]. Это предполагает систематизацию, обеспечивающую качество и полноту изложения. Использование кодов обеспечивает независимость интерпретации от субъективных факторов и позволяет применять автоматизированные (компьютерные) методы для обработки таких документов.

В данной статье рассматриваются современ -ные средства и форматы структурного представления и обработки медицинской информации.

История вопроса

Методы структурирования данных в сфере информационных технологий развивались вместе с ростом возможностей вычислительной техники. В тех приложениях, где данные

имеют относительно простую структуру, широко используются реляционные СУБД, а информационные системы часто имеют двух-звенную клиент-серверную архитектуру. Однако сфера медицинской информатики, где часто приходится иметь дело с изображениями и большими массивами свободного текста, не очень хорошо вписывается в рамки реляционной модели Кодда. Появляющиеся в последнее время объектно-ориентированные СУБД, вероятно, смогут в большей степени помочь решению возникших проблем, однако такие системы пока недостаточно стандартизированы и весьма дороги.

Ещё одной проблемой в медицинской информатике является необходимость обеспечения достаточно высокого уровня интеллектуальности информационных систем. Обычно из этого вытекает потребность в специфическом представлении информации, то есть в особых форматах данных, которые отсутствуют в традиционных СУБД.

Известно, что узкоспециальные решения вышеобозначенных проблем существуют уже достаточно давно, однако широкое распространение сети Интернет требует глобальной стандартизации средств представления, обработки и передачи данных. Данная статья - обзор современных решений в этой области, а также попытка обозначить их сильные и слабые стороны с учётом специфики медицинской информатики.

SGML

SGML - это общепризнанный международный стандарт (ISO 8879-1986), интенсивно используемый, например, правительством США во многих, в том числе и негражданских проектах. SGML - это метаязык (т.е. средство формального описания языков), который определяет правила создания грамматики документов. Формальная спецификация разметки для каждого типа документов обозначается термином Определение Типа Документа (Document Type Definition - DTD). DTD определяет структуру, которую будут иметь все документы данного типа. В DTD определены все теги (tag), т.е. элементы структуры документа, а также отношения между элементами данных. Это значит, что любой документ известного типа может быть стандартно обработан известным способом. SGML (путём разметки) сводит документ или сообщение к набору понятий в заданной грамматике.

Каждый элемент отмечается начальным тегом и, как правило, конечным тегом. Например, элемент Диагноз в тексте может быть отмечен так:

Ангина

Элемент может иметь атрибуты, которые несут дополнительную информацию о нем. Атрибуты позволяют использовать один элемент во многих контекстах вместо того, чтобь создавать множество похожих друг на друга элементов. Имена и значения атрибутов записываются в начальном теге элемента и отделяются друг от друга знаком равенства.

Объекты (entities) представляют собой мощное и гибкое средство адаптации стандартного DTD-определения к специфические приложениям (например, к кардиологии) или подразделам этих приложений (например, катетеризация). Для построения жизнеспособной инфраструктуры медицинской документации необходимо подготовить набор специфических SGML-тегов (общих идентификаторов) и других SGML-"фрагментов", которые затем могли бы использоваться во многих контекстах.

Для совместного анализа SGML-документа и его DTD с целью определения правильности данного документа (т.е. соответствует ли он правилам, заложенным в DTD), компьютер использует специальное ПО, называемое анализатором или парсером (parser). Парсер требуется также и на этапе составления и модификации DTD-определений для проверки их структурной корректности. Парсер может проверить, присутствуют ли в документе все необходимые элементы и следуют ли они в правильном порядке.

В связи с тем, что SGML работает с фрагментами структурированных документов, имеется возможность создания целых документов из частей, предоставляемых различными подразделениями организации.

Важно отметить, что SGML не определяет собственно структуру документа или семантику тегов - вместо этого он определяет способы их задания. В SGML есть определённые типы разметки (и, соответственно, группы тегов):

- процедурная разметка, определяющая представление (внешний вид) данных. Например, тег **задаёт утолщённый (bold) шрифт.**
- описательная (или семантическая) разметка, задающая смысл, структуру расположения данных.

SGML в медицинской информатике

Сегодня насчитывается четыре основных области применения SGML в медицинской информатике: медицинские публикации, представление новых лекарств, руководства по медицинской практике и электронные медкарты пациентов.

SGML используется для составления руководств по медицинской практике в Национальной Медицинской Библиотеке (National Library of Medicine) США. Использование SGML в электронных медкартах пациентов находится в стадии экспериментальных разработок. В 1996 году инициативы группы HL7 в области работ с SGML привели к созданию исследовательской подгруппы (получившей кодовое наименование HL7 SGML SIG) в рамках проекта HL7 [2, 3, 4]. Эта подгруппа занята разработкой исчерпывающей архитектуры медицинских документов, развитием и поддержанием всей инфраструктуры медицинской документации. Архитектура, разрабатываемая подгруппой HL7 SGML SIG, будет согласована с моделью данных, разрабатываемой группой HL7.

По причине того, что SGML имеет опциональные (необязательные) поля, написание парсеров и браузеров для него является трудной и дорогой задачей. По этой причине SGML не нашёл пока широкого применения в меди-динской информатике.

HTML

HTML - это производный от SGML язык, ориентированный в первую очередь на компоновку Web-документов. Сегодня HTML рассматривается как универсальный способ отображения данных.

HTML имеет своё специфическое DTD-определение, задающее процедурную разметку например, теги **- утолщённый шрифт, - большой шрифт для заголовка,...), некоторые элементы описательной разметки - заголовок документа, -тело документа,...) и гипертекстовую разметку (- ссылка).**

Хотя стандарты пользовательского интерфейса и отображения данных являются необходимым элементом при работе с информацией, они не определяют достаточных средств доступа к самим текстовым и графическим данным. В частности, они не задают механизмов интеллектуального поиска, передачи, адаптируемого представления и прочих манипуляций с информацией в разнообразных контекстах.

HTML не может удовлетворить эти потребности, т.к. он является языком, описывающим то, как Web-страница должна выглядеть, а не то, как

организовывать и описывать данные.

Резюмируя, скажем, что хотя HTML и имеет широкие возможности по отображению информации, он не имеет серьёзных средств управления этой иформацией. Поэтому он не является интересным с точки зрения предмета настоящей статьи, и в дальнейшем рассматриваться не будет.

XML

XML - это текстовый формат, позволяющий разработчикам описывать данные, представлять и хранить их в структурированном виде. XML также позволяет передавать структурированную информацию между информационными системами, обеспечивая при этом единообразие передаваемых данных и независимость их от обрабатывающих приложений и операционных сред.

Спецификация XML версии 1.0 была представлена в феврале 1998 года консорциумом World Wide Web Consortium (W3C). В число участников рабочей группы, разрабатывающей стандарт XML, входят такие крупные производители ПО как Microsoft, Netscape Communications, Sun Microsystems и другие [5].

XML (в отличие от HTML), позволяет определять неограниченное множество тегов, что даёт разработчикам возможность гибко выбирать, какие данные и каким образом использовать.

Обзор, синтаксис и преимущества XML

Синтаксис XML прост:

содержимое элемента

Закрывающий тег (отмеченный знаком /) является обязательным, и пересечение тегов не допускается. Если между тегами ничего нет, то тег может быть один:

Использование XML даёт разработчикам и пользователям ряд существенных преимуществ:

- Более удобный поиск.
- Возможность разработки гибких Web-приложений.
- Интеграция данных из разнородных источников информации.
- Интеграция данных от различных приложений.
- Возможность локальных вычислений и манипуляций данными (т.е. комбинация локальных и удалённых вычислений).
- Множественные представления одних и тех же данных.
- Поддержка частичных обновлений (в этом случае требуется обновить только ту часть информации, которая была изменена - отсюда и меньший сетевой трафик).

- Открытые и развивающиеся стандарты (к которым и принадлежит семейство стандартов XML).
- Представление информации в Интернете (XML - это текстовый формат, который может быть использован при передаче по HTTP точно так же, как это делается и с HTML).
- Улучшенная масштабируемость приложений (т.е., опираясь на XML, разработчики могут внедрять в документы особые процедурные описания, содержащие информацию о том, как обрабатывать данные того или иного типа).

- Поддержка компрессии (XML- документы компрессируются очень хорошо из-за частой повторяемости тегов).
- Поддержка со стороны ведущих производителей ПО (Morosoft, Sun, Grade и др.)

Любой XML-документ может дополняться его структурным определением (см. выше об аналогичных принципах в SGML) - DTD (Document Type Definition). В DTD определяются требования к структуре документа. DTD-определения позволяют проверять корректность данных в случае, когда приложение-получатель не имеет встроенного описания входящей информации. Однако наличие DTD является необязательным. Данные, переданные вместе с DTD, обозначаются термином "истинный" ("valid") XML. В этом случае XML-парсер может проверить структурную корректность входящей информации в соответствии с правилами, определёнными в полученном DTD. Данные, переданные без DTD, обозначаются (конечно, только в случае корректности самого XML-документа) термином "правильно построенный" ("well-formed") XML.

Открытость и гибкость стандарта XML позволяет использовать его в любой области, где есть потребность в обмене и передаче информации. Хорошо организованное DTD-определение гарантирует надёжность обмена документами и целостность передаваемой информации.

Трёхзвенная архитектура. Интеграция данных.

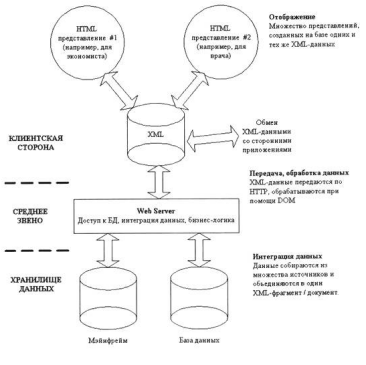
При использовании масштабируемой трёхзвенной архитектуры XML-документы могут быть автоматически сгенерированы из существующих баз данных. XML позволяет отделить данные от правил их обработки и от механизмов отображения. Интеграция, передача, обработка и отображение данных - вот основные этапы манипулирования информацией [6]:

Интеграция данных.

Приложения, выполняющиеся на среднем звене, имеют доступ к множеству баз данных и могут переводить существующую в них информацию в XML.

Компания Microsoft предложила концепцию схемы данных XML (XML Data schema) для улучшения интеграции с реляционными СУБД [7]. Схема - это формальная специфика-ция наименований элементов, которая определяет, какие элементы, и в каком порядке могут встречаться в XML-документе, какие они могут иметь атрибуты и дочерние элементы. В этом смысле схема - это функциональный эквивалент DTD. Однако, в отличие от DTD, схемы данных XML сами написаны на XML, что делает XML самоопределяющимся языком. Также схемы дают разработчику такие средства, как типизация данных, наследование и определение правил представления. С этой точки зрения схемы являются весьма выгодной альтернативой DTD-определениям, гарантирующей улучшенную интеграцию данных, хранимых в реляционных базах данных.

Используя возможности наследования, разработчики могут создавать новые специфические схемы на основе уже существующих схем общего назначения. Например, можно создать схему для отчёта ультразвукового исследования на основе общей схемы лабораторного отчёта.



"Трёхзвенная архитектура информационной системы"

Словари XML, среда определения ресурсов.

Словари XML - это наборы реальных понятий, используемых в конкретных типах XML-документов. Словари и структурные взаимоотношения между элементами могут быть формально определены в схемах данных XML или DTD-определениях как форматы данных, используемые в определённых типах документов. В среде интранет (например, в пределах учреждения здравоохранения) можно легко и быстро создать XML-словари с целью организации передачи данных в формате XML между приложениями. И в ходе эксплуатации такой системы можно без труда подстраивать описания передаваемых данных под возникающие потребности ЛПУ. В среде Интернет потребуется согласование различных локальных словарей с целью обеспечения возможности обмена информацией между разными организациями. В настоящее время в мире идут подобные разработки. Результатом одной из них является инфраструктура RDF.

Среда определения ресурсов (Resource Definition Framework, RDF) - это основа для обработки метаданных, обеспечивающая интероперабельность (т.е. возможность обмена данными) между приложениями, передающими машинно-читаемую информацию через Web. RDF предоставляет возможности автоматической работы с Web-ресурсами.

RDF, в отличие от XML - не формат и не язык, а скорее интерфейс (API). При помощи DOM (Document Object Model - Объектная модель документа, это своего рода интерфейс к XML-документам [8]) программист может получить доступ к физической структуре документа как к набору узлов, ссылок, элементов, атрибутов и т.д. Это удобно для работы с документом на физическом уровне, но не позволяет понять логическую структуру документа, смысл отдельных XML-тегов и суть их взаимоотношений с другими тегами.

Для решения этих проблем и создан RDF. Эта новая объектная модель (тоже своего рода API) даёт разработчикам доступ к логическому смыслу содержимого XML-документов. При помощи внешних файлов (называемых "схемами"), определяющих логический смысл и отношения для каждого XML-элемента, разработчики могут строить системы, в которых средства записи и чтения данных (для этого используется XML) отделены от средств интерпретации этих данных (для этого предназначены указанные "схемы"). С помощью такого подхода разработчики могут сконцентрироваться на бизнес-логике и понятиях конкретной предметной области, а не на технических деталях реализации системы.

XML (посредством DOM) обеспечивает возможность просмотра физической структуры документов. RDF, являющийся средством более высокого уровня, располагается "над" XML, обеспечивая возможность просмотра логической структуры документов. То есть можно сказать, что RDF - это язык описания логической семантики документа. Подробнее RDF описан в отчётах консорциума W3C (W3C Metadata activity files).

PICS

PICS - это специфический стандарт описания содержимого документов, предназначенный для выставления рейтингов Web-страницам и документам вообще. Этот стандарт может оказать неоценимую помощь в организации многоуровневого медицинского документооборота, в котором документы могут ранжироваться по многочисленным параметрам, например, таким, как срочность (cito !), важность, полнота, интересность и т.п.

Таким образом, подводя итог всей статьи, можно предложить следующую архитектуру современной многоуровневой медицинской информационной системы: в качестве базы используется трёхзвенная архитектура, описанная выше, данные в системе физически представляются в формате XML, над XML располагается и работает RDF, а над RDF - инфраструктура PICS.

ЛИТЕРАТУРА

1. **N. Balogh - Structured data transmission for CEC/NIS cardiology - INCO-COPERNICUS, 1999.**
2. **Andrew Hinchley, Short Strategic Study: Enabling Technologies - SGML/XML. CEN/TC 251/N98-061, 1998.**
3. **Robert H. Dolin, Liora Alschuler, Tim Bray, John E. Mattison, SGML as a Message Inter-change Format in Healthcare, JAMIA, 1997.**
4. **Liora Alschuler e. al, (Kona Editorial Group), Patient Record Architecture HL7 Document HL7 SGML/XML SIG September 4, 1998.**
5. **Tim Bray, Jean Paoli, C M. Sperberg-McQueen, Extensible Markup Language (XML) 1.0 W3C Recommendation 10-February-1998, http://www.w3.org/TR/REC-xml .**
6. **Joachim Dudek, David Markwell: XML Task Force -progress report to TC251.**
7. **Microsoft XML SBN Workshop, http://www.microsoft.com/xml .**
8. **W3C DOM WG, The Document Object Model (DOM) Level 1 Specification, W3C Recom-mendation, October 1, 1998. http://www.w3.org/TR/REC-DOM-Level-1 .**

Статья впервые опубликована в журнале "Информационные технологии в здравоохранении", 3-4 за 2003 год и воспроизводится с разрешения автора.