Функциональные методы обработки слабоструктурированных данных и их применение для построения электронных библиотек.

Журавлева О.В., Лисовский К.Ю. - Московский Институт Стали и Сплавов, Томусяк Г.С. – Л.А.С.П., Томусяк Э.С. - АВЭР

Значение XML как единого стандарта обмена данными можно сравнить с ролью Java для переносимого программного обеспечения. В настоящий момент "XML-революция" может считаться сосоявшимся фактом, а стремительное развитие XML-технологий в течении последних лет затронуло практически все отрасли информатики.

Электронные библиотеки не являются здесь исключением, и одним из важных преимуществ применения XML является упрощение интеграции электронных библиотек с Web.

В данном докладе мы рассматриваем простой для реализации метод обработки XML-метаданных электронной библиотеки. Подобные функциональные методы построения информационных систем Internet/intranet/extranet активно исследуются и используются [6],[8].

Применение их для работы с метаданными электронной библиотеки может существенно упростить построение интегрированной web-доступной информационной системы, так как именно метаданные используются обычно для поиска информации в электронных библиотеках.

XML Infoset [17] - это абстрактный набор данных, который описывает информацию находящуюся в корректно сформированном (well-formed) XML документе. Собственно язык XML предоставляет собой синтаксис для текстового описания структурированных данных, и является реализацией XML-Infoset.

Хорошо известно что для представления данных иерархической структуры могут быть эффективно использованы вложенные списки. Языки семейства Лисп испольуют для этой цели S-выражения.
S-выражение может быть рекурсивно определено как список, элементами которого являются либо атомарные значения, либо другие S-выражения.
Их возможности по представлению слабоструктурированных данных таковы, что позволяют говорить о XML как о нотации для записи S-выражений [10]. Соответственно, возможна и реализация XML-Infoset при помощи S-выражений, SXML [14].

Существуют реализации стека SXML технологий, являющихся аналогами XPath, XSLT, DOM, SAX, etc. Эти реализации используют функциональный язык программирования Scheme [6], один из диалектов Лисп, и предоставляют средства обработки SXML данных на Scheme.

Рассмотрим упрощенный пример представления католожной записи.
XML:

```
<?xml version="1.0"?>
<book>
<file>pushkin.ps</file>
<title>Избранное</title>
```

```
<author>A.C. Пушкин</author>
<publisher>"Просвещение"</publisher>
<address><city>Москва</city></address>
<year>1997</year>
</book>
```

SXML:

```
(book
 (file "pushkin.ps")
 (title "Избранное")
 (author "A.C. Пушкин")
 (publisher "Просвещение")
 (address
   (city "Москва"))
 (year "1997"))
```

Отметим, что SXML данные являются в то же время и синтаксически корректной программой Scheme, причем символы тегов выполняют роль функций применяемых к содержимому соответствующих XML элементов. Связав эти тэги с функциями обработки соответствующих элементов возможно осуществить преобразование SXML данных путем непосредственного выполнения их как функциональной программы [7].
Синтаксис Scheme, как и других языков семейства Лисп, основан на представлени программ в виде S-выражений, и стандарт языка Scheme [13], предоставляет функцию `eval`, вполняющую S-выражение как программу.

Наиболее очевидным является применение такого подхода для генерации HTML, однако он может быть использован и для других задач, таких как генерация форм запросов к каталогу библиотеки, генерация печатных отчетов, организация репликаций etc.

При наличии DTD или XML-Schema они могут использоваться для автоматической генерации функциональных связываний. Сочетание такой технологии с DataGuide позволяет работать с "каталогами" не имеющими предопределенной фиксированной структуры.
Эти возможности представляются особенно ценными для создания "легковесных" электронных библиотек или коллекций, которые создаются и поддерживаются не специализированными организациями.

Существенно повысить гибкость обработки XML данных позволяет использование языков запросов. Требования к такому языку сформулированны в XML Query Requirements[10] предложенной W3C XML Query Working Group. В качестве прототипа для реализации на SXML был выбран QUILT [3]. Этот язык вобрал в себя лучшие черты своих предшественников. Из XPath[15] XQL[9] взят синтаксис навигации по иерархичесим документам. Из XML-QL[5] связывание переменных и их использование для создания новых структур. Из SQL[12] пришла идея использования последовательности ключевых слов, шаблону SELECT FROM WHERE сопоставлен шаблон FOR LET WHERE RETURN (FLWR).

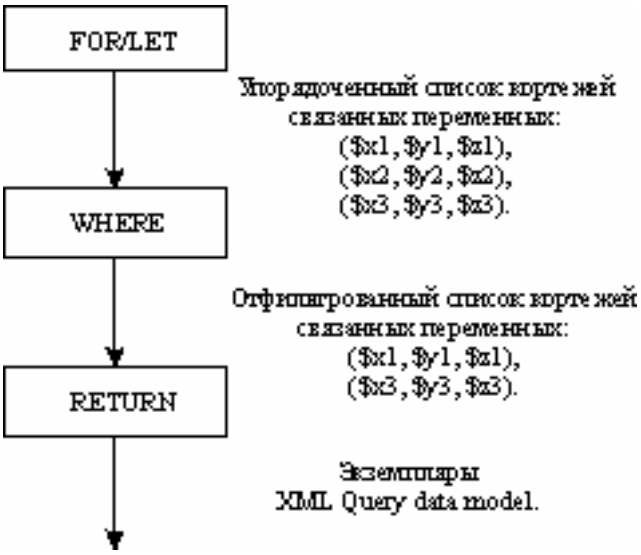

Рис.1. Поток данных в FLWR

Подобно тому как язык SQL является замкнутым на множестве реляционных отношения, QUILT является замкнутым относительно документа, фрагмента документа или коллекции документов, то есть данных соотвествующих XML Query Data Model[16].

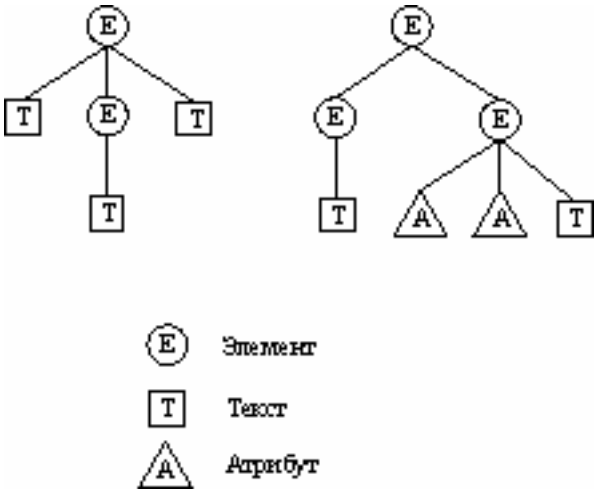

Рис.2. Пример XML Query Data Model: упорядоченный лес деревьев

Повлияли на синтаксис QUILT и другие языки запросов такие как: OQL[2], Lorel[1], YATL[4] и др. Важным преимуществом языка QUILT перед его конкурентами является более высокая степень соответствия требованиям изложенным в [18]. По своему построению Quilt является функциональным языком, что упрощает его реализацию в среде SXML/Scheme.
В настоящий момент мы располагаем работающим прототипом функциональной реализации подмножества языка Quilt.

`Quilt` запрос представляется в следующем виде.

```
((for (book)
    "library.xml")
```

```
(where (and
        (= (address city *data*) "Москва")
        (> (year *data*) 1997)))
(return
  (book
    (@xpath (author))
    (@xpath (title))
    )))
```

В Московском Институте Стали и Сплавов рассматриваемый подход применяется в течении полутора лет для построения электронной библиотеки научно-технических публикаций и отчетов. Система поддерживает добавление и изъятие документов, изменение и модификацию метаданных, и поиск по метаданным стандарта DublineCore [11]. Накопленный опыт позволяет говорить о простоте реализации и сопровождения легковесных электронных коллекций и библиотек с использованием функционального подхода.

Список литературы:

[1] Serge Abiteboul, Dallan Quass, Jason McHugh, Jennifer Widom, and Janet L. Wien: The Lorel Query Language for Semistructured Data.
International Journal on Digital Li-braries, l(1):68-88, April 1997.

[2] Rick Cattell et al. The Object Database Standard. ODMG-93, Release 1.2. Morgan Kaufmann Publishers, San Francisco, 1996.

[3] D. Chamberlin, J. Robie, and D. Florescu Quilt: An XML Query Language for Heterogeneous Data Sources
// Lecture Notes in Computer Science, Springer-Verlag, 2000.

[4] S. Cluet, S. Jacqmin, and J. Simeon: The New YATL. Design and Specifications. Technical Report, INRIA, 1999.

[5] Alin Deutsch, Mary Fernandez, Daniela Florescu, Alon Levy, and Dan Suciu: A Query Language for XML.
http://www.research.att.coml-mff/files/final.html

[6] FramerD: Representing Knowledge in the Large
IBM Systems Journal (Volume 35, Numbers 3-4, 1996)

[7] Oleg Kiselev. XML and Scheme
Workshop on Scheme and Functional Programming 2000, Montreal, 2000

[8] Kurt Normark. "Programming World Wide Web pages in Scheme" Sigplan Notices, vol. 34, no. 12, 1999

[9] J. Robie, J. Lapp, D. Schach: XML Query Language (XQL).

[10] P. Wadler. The Next 700 Markup Languages
// Second Conference on Domain Specific Languages (DSL'99), Austin, Texas, October 1999.

[11] Dublin Core http://dublincore.org/

[12] International Organization for Standardization (ISO).
Information Technology-Database Language SQL. Standard No. ISO/IEC 9075:1999.
(Available from American National Standards Institute, New York, NY 10036, (212) 642-4900)

[13] Revised (5) Report on the Algorithmic Language Scheme
Revised5 Report on the Algorithmic Language Scheme, R. Kelsey, W. Clinger, J. Rees (eds.),
Higher-Order and Symbolic Computation, Vol. 11, No. 1, September, 1998

[14] SXML Specification http://zowie.metnet.navy.mil/~oleg/ftp/Scheme/SXML.html.

[15] World Wide Web Consortium. XML Path Language (XPath) Version 1.0. http://www.w3.org/TR1xpath.html
W3C Recom-mendation, Nov. 16, 1999.

[16] World Wide Web Consortium. XML Query Data Model. http://www.w3.org/TR/query-dsatamodel
W3C Working Draft, May 11, 2000.

[17] XML Information Set http://www.w3.org/TR/xml-infoset/
W3C Working Draft 2 February 2001

[18] XML Query Requirements. http://www.w3.org/TR/xmlquery-req
W3C Working Draft 15 February 2001

FUNCTIONAL APPROACH TO SEMISTRUCTURED DATA MANAGEMENT AND ITS APPLICATION IN IMPLEMENTATION OF DIGITAL LIBRARIES

O. Juravleva, K. Lisovsky, E. Tomusjak, G. Tomusjak

This presentation considers functional programming techniques of semistructured data management and their application to digital libraries. Benefits of S-expressions for representation of XML data and possible ways of their utilization are considered.

The presentation also considers the role of the semistructured data schema in context of the functional programming approach and proposes a technique for implementation of digital libraries and collections without a predefined schema.

Possibility for implementation of a Quilt-like query language is discussed and demonstrated.