АЛГОРИТМИЗАЦИЯ И ПРОГРАММИРОВАНИЕ

УДК 004.4

Е. В. Кабак, ст. преподаватель (БГТУ); Н. П. Коровкина, доцент (БГТУ);
Н. Н. Пустовалова, доцент (БГТУ)

МОДЕЛИ И ЯЗЫКИ ОПИСАНИЯ ТЕХНИЧЕСКИХ ЗНАНИЙ

Статья содержит обзор существующих языков специализированной разметки, которые используются для описания знаний из различных естественно-научных и технических областей. Особое внимание уделено языкам группы «X-технологий», таким как CML, MathML, GML, SVG, SMIL, MatML, CellML, PML, X3D и др. Рассмотрены преимущества и недостатки перечисленных средств. Проанализированы программные средства, которые используются для работы с документами специализированной разметки. Затронуты также проблемы эффективной обработки, оперативной актуализации, интеграции, многократного и многоцелевого использования фрагментов таких документов в различных контекстах; назначение и особенности новых Интернет-технологий и их роль в создании эффективного и качественного контента.

Article contains the review of existing languages of a specialized marking which are used for the description of knowledge from various natural-science and technical fields. The special attention is given languages of «X-technologies» group, such as CML, MathML, GML, SVG, SMIL, MatML, CellML, PML, X3D, etc. Advantages and disadvantages of the listed languages are considered. Software which is used for process such documents of a specialized marking is analyzed. Problems of effective processing, operative actualization, integration, reusability and multi-purpose multiple uses of fragments of such documents in various contexts; appointment and features of new Internet technologies and their role in creation of an effective and qualitative content are considered also.

Введение. Область технических знаний в современных условиях предъявляет к языкам описания структуры и содержимого электронных документов большое число требований. Современные языки разметки должны облегчить обработку, интегрирование данных разного рода, что предусматривает необходимость создания конечного набора соглашений, правил разметки, которые бы обеспечили потенциал взаимосвязи электронных сред, а также способствовали бы преодолению трудностей, связанных с представлением технических знаний на экране. Кроме того, при выборе языков описания структуры и содержимого технических документов, при разработке специализированного программного обеспечения необходимо учитывать возможность решения с их помощью более важных задач, чем простое отображение документа. Здесь становится необходимым найти способы и средства раскрыть смысл документа для того, чтобы облегчить его автоматическую обработку, поиск, индексацию и одновременно обеспечить его многоцелевое и многократное использование в других контекстах.

Еще в конце 60-х гг. XX в. сотрудники IBM Ч. Гольдфарб, Э. Мошер, Р. Лори сформулировали три общих принципа, которые гарантируют взаимодействие между программами, выполняющими обработку документов [1]:

1) использование единых принципов форматирования, т. е. наличие единого набора синтаксических конструкций и общей семантики;

2) специализация языков форматирования;

3) четкое определение структурного формата документа, что предполагает формулировку правил, определяющих количество и маркировку языковых конструкций, используемых в документе.

Следование этим принципам в первую очередь означает возможность отделения контента документа (его информационного наполнения) от разметки для эффективного выполнения процедур обработки, поиска, анализа и представления информации на более высоком технологическом уровне [2]. При этом значительный упор делается в сторону машинной обработки информационных ресурсов.

Сегодня также актуальным становится вопрос о необходимости соответствия документов современных информационных ресурсов, наряду с уже перечисленными принципами, следующим основным критериям [2]:

– сохранение способности быть прочитанным и интерпретированным в течение

многих лет независимо от изменений технических, программных и других средств их обработки;

– обеспечение доступности информации для выделения и обработки с помощью различных технологий.

Основная часть. В случае соответствия документа перечисленным критериям и принципам допустимо утверждать, что он будет находиться в технологически безопасной форме, способной обеспечить эффективную работу с документом с точки зрения его многоцелевого, многократного использования.

Основу достижения поставленных целей составляет стандарт XML и связанные с ним так называемые «X-технологии» [3], которым в современных системах документооборота, обработки и хранения информации отводится роль ключевых технологий [3–5]. Практически все языки специализированной разметки, которые рассматриваются в данной статье, созданы на основе метаязыка XML: MathML, CML, GML, SMIL, SVG, X3D и др.

Стандартизация и унификация языков описания данных и знаний, следование компаний-разработчиков программного обеспечения международным спецификациям и рекомендациям позволяют свести к минимуму собственные проектные усилия, интегрировать в единую целостную систему различные приложения, создавать информационные ресурсы, различные компоненты которых можно обрабатывать различными программами, на различных аппаратно-программных и технологических платформах, в разных информационно-технологических системах [2].

В настоящее время большинство технических электронных документов представляют собой совокупность текстов, приемов их форматирования, а также изображений, чаще всего сохраненных в формате JPG, BMP, PNG. Широко используемым является также формат PDF. Эти форматы представления технических знаний не всегда характеризуются высоким качеством отображения информации, являются довольно примитивными и неадекватны современным требованиям науки. За рубежом технические знания часто описываются языком T_EX (Д. Кнут) и отображаются с помощью специальных программных пакетов, например LaTeX, которые практически не используется на постсоветском пространстве [1].

Перечислим наибольшие проблемы, связанные с традиционными способами хранения, представления и передачи технических текстов:

1) проблема отображения, включая вывод на экран и на твердый носитель;

2) проблема обработки: поиск, индексирование, выделение, масштабирование без потери качества, разделение на фрагменты с возможностью их использования в других контекстах или приложениях;

3) проблема передачи информации, например, в сети Интернет с учетом таких параметров, как время и стоимость. В частности, разметка описывает математическое выражение в более сжатом и кратком виде, чем, например, изображение.

Преимущества и достигаемые цели современных специализированных языков описания технических знаний:

– представление материалов в виде электронных документов, доступных для прочтения человеком (с помощью и без помощи специальных программ);

– объединение в единое целое структуры некоторого выражения и его значения;

– конвертирование материалов в различные широко используемые форматы с помощью различных инструментальных программных средств;

– хранение информации в форме, обеспечивающей возможность ее различных вариантов представления и использования;

– наглядное представление технических знаний с возможностью масштабирования отдельных фрагментов документа без потери качества;

– обеспечение возможности широкого распространения с минимальными затратами, использования материалов, их редактирования, дополнения и т. д.;

– создание шаблонов и других документов, облегчающих редактирование и представление технических материалов;

– обеспечение интерактивности отдельных фрагментов технического документа при необходимости и др.

Минимальные функциональные возможности, которые должно предоставлять инструментальное средство создания и распространения технических материалов, статей в этой связи:

1) технические знания должны отображаться должным образом, в соответствии с предпочтениями читателей и авторов материалов и с учетом максимальных возможностей, предоставляемых конкретной технологией;

2) технические документы, содержащие, например, математические выражения, химические формулы, технические объекты, должны выводиться на печать должным образом с учетом разрешения принтера;

3) фрагменты электронных технических документов должны быть в состоянии реагиро-

вать на действия пользователя, в том числе для того, чтобы обеспечить взаимодействие с другими программными средствами;

4) редакторы технических электронных документов должны быть развитыми, гибкими, расширяемыми программными средствами, подходящими для обеспечения взаимодействия и интеграции программного обеспечения, способными к созданию высококачественных материалов для их представления и размещения в сети Интернет;

5) редакторы технических электронных документов должны предоставлять простые способы создания технической документации, автоматической генерации кода документов, возможности их редактирования вручную;

6) авторы и читатели должны быть абсолютно свободными в выборе средств создания и просмотра технических документов.

Перейдем к конкретному рассмотрению существующих специализированных языков описания естественно-научных и технических знаний.

В первую очередь рассмотрим языки математической разметки. Математическая система обозначений имеет сложную структуру, состоящую из математических формул, выражений и текста. Основные проблемы, возникающие при записи математических выражений и формул, можно разделить на две группы: проблемы кодирования и проблемы реализации. Проблемы, связанные с включением в документ математических записей, относятся к проблемам реализации. Проблемы, связанные с автоматической обработкой данных, относятся к проблемам кодирования, поскольку математические записи, выражения более трудны для обработки, чем обычный текст. Обе из перечисленных проблем достаточно эффективно решает язык математической разметки MathML.

Язык *MathML* (*Mathematical Markup Language*) – язык описания математических формул и выражений с использованием синтаксиса XML. В настоящее время язык MathML фактически стал стандартом представления математической информации в электронной форме. Синтаксис языка MathML довольно прост. Приведем в качестве примера (рис. 1) фрагмент документа MathML (расширение MML), содержащего формулу

$$4x^2 - 5x + 6 = 0.$$

К программным средствам обработки документов MathML относится браузер Amaya, разработанный консорциумом W3C. Пример оформления математической статьи в редакторе Amaya показан на рис. 2.

```
<math xmlns="http://www.w3.org/1998/Math/ MathML">
  <mn>4</mn>
  <msup>
    <mi>x</mi>
    <mn>2</mn>
  </msup>
  <mo>&minus;</mo>
  <mn>5</mn>
  <mi>x</mi>
  <mo>+</mo>
  <mn>6</mn>
  <mo>=</mo>
  <mn>0</mn>
</math>
```

Рис. 1. Организация математической информации в файле формата MML

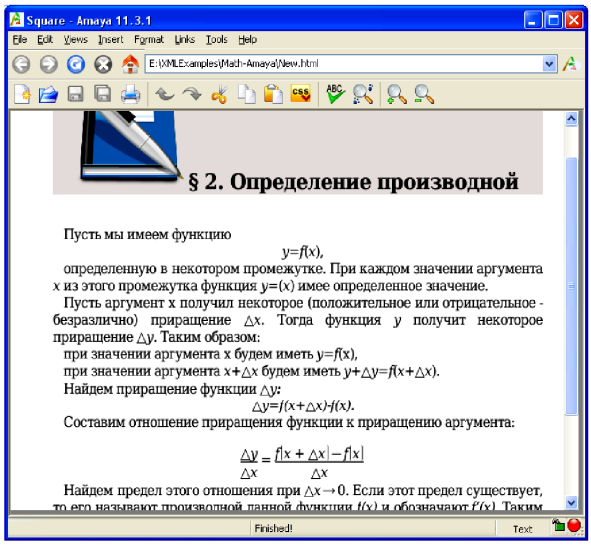

Рис. 2. Фрагмент математического документа, созданный в редакторе Amaya

Химический язык разметки *CML* (*Chemical Markup Language*) является аналогом MathML для создания материалов по химической тематике. Язык CML представляет собой язык на основе XML (также использует Java) для описания молекулярных структур, записи химических формул, химических реакций, обработки данных о химических соединениях, разработанный как часть проекта *Open Molecule Foundation*. Данный язык способен поддерживать крайне сложные информационные структуры. Близкими к CML по назначению являются языки, разработанные впоследствии: Analytical Information Markup Language (AniML), Bioinformatic Sequence Markup Language (BSML), BIOpolymer Markup Language (BIOML), CellML, Computational Chemistry Markup Language (CCML), SpectroML, ThermoML и ряд других. К программным средствам обработки CML относятся: Jumbo 3, ChiMeraL, Jmol, Marvin (ChemAxon) и т. п. К программным средствам обработки CML относятся: Jumbo 3, ChiMeral, Jmol, Marvin (ChemAxon) и др.

Фрагмент химического документа, созданного с помощью редактора Marvin, показан на рис. 3. Трехмерное представление этого же фрагмента приведено на рис. 4.

Назначение языка *CellML* (*Cell Markup Language*) заключается в электронном хранении и обмене математическими моделями. Данный язык широко применяется в биологическом моделировании, поддерживает спецификацию MathML.

Язык *MatML* (*Materials Markup Language*) предназначен для описания свойств материалов.

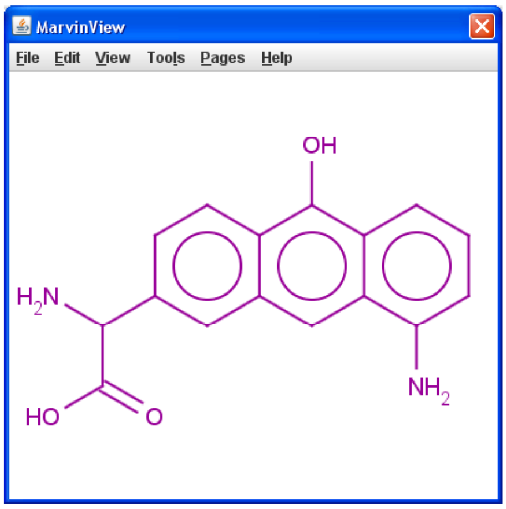

Рис. 3. Фрагмент документа, созданного на основе языка CML

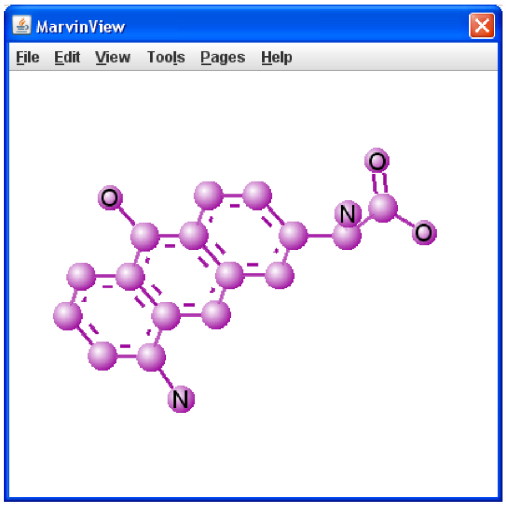

Рис. 4. Трехмерное представление документа в редакторе Marvin

Язык *GML* (*Geography Markup Language*) используется географическим сообществом. В GML географическая информация описывается с помощью объектов местности. Каждый объект местности включает в себя свойства и геометрии. Объекты местности, по сути, представляют собой геометрические примитивы. Геометрии содержат географические данные, а свойства добавляют к этим данным описательную информацию.

PML (*Physical Markup Language*) – язык общего назначения для описания физических объектов и сред для промышленных, коммерческих и потребительских приложений. PML поддерживает такую модульность и гибкость, что его можно использовать при мониторинге и управлении физической средой. К числу приложений относится контроль состояния склада, автоматические транзакции, управление цепочкой поставки, машинный контроль и взаимодействие между объектами.

SMIL (*Synchronized Multimedia Integration Language*) дает возможность авторам документов координировать отображение различных мультимедийных элементов. В SMIL мультимедийные элементы могут работать совместно.

Язык разметки *SVG* (*Scalable Vector Graphics*) представляет собой один из способов описания данных двумерной векторной и смешанной векторно-растровой графики при ее использовании в Web.

Язык *X3D* (*Extensible 3D*) рекомендован консорциумом Web3D Consortium для использования в сети Интернет в качестве файлового формата, описывающего интерактивные 3D-объекты и миры.

Язык *STTML* (*Scientific Technical and Medical Markup Language*) служит основой для создания других специальных языков разметки, описания технической, научной, медицинской информации.

Все из перечисленных языков соответствуют новой идее развития так называемого семантического Интернета (семантического веба). Суть этого явления заключается в переводе документов сети из стандартного формата HTML в формат XML, что позволит компьютерам понимать смысл семантических данных документов.

Заключение. Таким образом, в данной статье выполнен обзор существующих в настоящее время специализированных языков разметки, позволяющих описывать содержимое документов из специальных областей знаний. В статье показаны новые возможности, которые получают авторы и читатели таких документов.

Следует отметить, что данная статья отражает промежуточные результаты исследования, посвященного изучению возможностей практического применения современных языковых и программных средств создания и представления естественно-научного и технического контента. Дальнейшими же и перспективными направлениями работ в обозначенной области являются:

– разработка общих рекомендаций для программного обеспечения, применяемого для создания, распространения, интегрирования технического электронного контента;

– определение требований к программному инструментарию для поддержки методики разработки технического электронного контента;

– разработка специального инструментального программного средства для создания, отображения технической и научной документации, статей, публикаций и т. д. из различных специальных областей знаний.

Литература

1. Как программировать на XML / Х. М. Дейтел [и др.]; пер. с англ. – 2-е изд. – М.: ООО «Бином-Пресс», 2008. – 944 с.

2. Курбацкий, А. Н. Построение ключевых элементов корпоративных информационных систем на основе XML-технологи / А. Н. Курбацкий, В. А. Чеушев, Бинь Сюе // Информатизация образования. – 2008. – № 4 (53). – С. 33–62.

3. Старыгин, А. А. XML. Разработка web-приложений / А. А. Старыгин. – СПб.: BHV-СПб, 2003. – 592 с.

4. Елизаров, А. М. Технологии управления разнородным естественнонаучным контентом на основе семантического веба / А. М. Елизаров, Е. К. Липачев, М. А. Малахальцев // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: труды 11-й Всерос. науч. конф., Петрозаводск, 9 сент. 2009 г. / НИИ математики и механики имени Н. Г. Чеботарева Казанского государственного университета. – Петрозаводск, 2009. – С. 325–328.

5. Елизаров, А. М. Языки разметки семантического веба. Практические аспекты: учеб.-метод. пособие по направлению «Электронные образовательные ресурсы» / А. М. Елизаров, Е. К. Липачёв, М. А. Малахальцев. – Казань: КГУ, 2008. – 64 с.

Поступила в редакцию 31.03.2010