Методы семантической разметки веб-документов

© К.А. Кудим

Институт программных систем НАН Украины, г. Киев

kuzmaka@mail.ru

Аннотация

Обсуждаются такие методы дополнительной семантической разметки внутри XHTML-страниц, как микроформаты, RDFa, eRDF и XML.

1 Введение

Последнее десятилетие ведутся активные работы по поиску путей структурирования массива информации в интернете. Основа этих работ – концепция семантической паутины [1], то есть идея создания такой сети, в которой данные были бы связаны не просто гиперссылками, но смысловыми отношениями. На основе новых стандартов предполагается создать всемирную семантическую сеть данных, предназначенную для эффективной автоматической обработки и существующую параллельно обычной гипертекстовой сети. Такая сеть связанных данных создается медленно, поскольку подавляющее большинство веб-разработчиков не заинтересовано в создании дополнительной функциональности с расчетом на неясные преимущества в будущем, так как реально полезных и достаточно распространенных веб-сервисов, которые использовали бы преимущества семантического представления данных, все еще нет. В то же время, сама полезность такого представления почти не вызывает сомнений. Проблема кроется в том, что семантическую паутину приходится создавать как бы заново, с нуля, параллельно существующей многонаполненной гипертекстовой сети. Один из эволюционных путей преодоления этого разрыва – постепенно делать разметку гипертекста все более приближенной к семантическому представлению. В какой-то мере такая семантическая разметка может быть осуществлена с помощью стандартных средств языка XHTML [2], например, при простом цитировании или создании простого списка терминов. Но его возможностей в настоящий момент явно недостаточно даже в самых простых случаях, как, например, нет специальных элементов и атрибутов для создания списка библиографии или описания человека. Хорошо развились два конкурирующих способа такой дополнительной семантической разметки – микроформаты [3] и RDFa [4], краткому рассмотрению которых и посвящена настоящая работа. Рассматривается также альтернативный способ внедрения семантических структур в гипертекст, основанный на расширении языка XHTML новыми элементами.

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

2 Семантическая разметка

Как было сказано во введении, современные стандарты XHTML [2] позволяют в определенной степени структурировать гипертекстовую информацию по смыслу.

Семантическая разметка (или семантическая верстка) – это метод разметки гипертекста, при котором выбор того или иного элемента языка разметки осуществляется не на основе предписанного ему способа отображения браузером, а на основе значения заключенных в нем данных.

Подмножество языка XHTML, отвечающее принципам семантической верстки, и соответственно размеченный текст называют семантическим XHTML, а также обозначают иногда аббревиатурой POSH (Plain Old Semantic HTML – простой старый семантический HTML).

Чтобы подчеркнуть отличия семантической разметки, перечислим несколько рекомендаций по использованию некоторых элементов XHTML:

- использовать элемент TABLE только для разметки таблиц, но не для форматирования взаимного расположения логически не связанных элементов страницы; для этой цели лучше подойдет элемент DIV или один из более специфичных элементов, подходящий по смыслу, например, ADDRESS для контактной информации и т. п.;
- исключить использование всевозможных элементов XHTML, предназначенных для изменения внешнего вида; не использовать изображения-заполнители для создания отступов и т. п.;
- исключить использование элемента BR для разделения блоков текста; для самих блоков использовать наиболее подходящий по смыслу элемент, например, P или H1;
- исключить использование элементов B и I для смыслового выделения; для этой цели лучше подойдут элементы EM и STRONG;
- использовать как можно более узкий по смыслу элемент, например, для разметки цитат следует предпочесть элементу DIV элементы Q и

BLOCKQUOTE в сочетании с CITE.

Использование подобных правил позволяет максимально реализовать возможности XHTML для создания семантической разметки без дополнительных средств.

3 Микроформаты

Выразительные возможности семантического XHTML ограничены небольшим набором элементов и атрибутов для часто встречающихся случаев разметки, в то время как существуют и другие часто используемые структуры данных, которым нет соответствия в XHTML. Для разметки некоторых таких структур применимы так называемые микроформаты [3].

Микроформат – это определенная модель данных и соответствующая структура разметки, использующая атрибуты XHTML, такие, как class, rel и rev, для выражения конкретного семантического значения размеченного блока. Микроформаты не расширяют язык XHTML, а лишь декларируют, какие значения атрибутов и какую структуру элементов следует использовать для того, чтобы разметка трактовалась определенным образом, и, наоборот, как интерпретировать такую разметку. Стоит отметить, что с точки зрения микроформатов атрибут class рассматривается не только как возможность обозначить стиль элемента, но и, более широко, как тип элемента вообще.

На уровне стабильных версий существуют следующие микроформаты:

- hCalendar – служит для разметки событий;
- hCard – описание людей и организаций;
- rel-license – обозначает, что ресурс по ссылке является лицензией для ссылающейся страницы;
- rel-nofollow – указание поисковой машине, что не следует учитывать эту ссылку для подсчета рейтинга страницы по ссылке (запрет перехода по ссылке);
- rel-tag – указывает, что данная ссылка является ссылкой на тематическую рубрику (тэг) для ссылающейся страницы или ее части;
- VoteLinks – указывает отношение автора ссылки к ресурсу по ссылке, используется одно из трех возможных значений: «за», «нейтрально», «против»;
- XFN – описание отношений между людьми с помощью гиперссылок;
- XMDP – микроформат для описания микроформатов;
- XOXO – описание произвольной XML структуры в рамках XHTML.

Еще 17 микроформатов находятся в состоянии черновиков, то есть спецификации для них уже хорошо проработаны, но еще не окончательны и могут подвергаться изменениям.

3.1 Примеры использования микроформатов

На рисунках 1 и 2 приведены примеры разметки с помощью микроформатов hCalendar и hCard. Из этих двух примеров можно видеть, что использование микроформатов не выходит за рамки обычной разметки, разве что для определения стилей количество классов избыточно.

На рис. 3 приведен пример использования микроформата rel-nofollow. Автоматическое добавление такого простого атрибута, например, при публикации комментариев посетителей ресурса, позволяет ограничить пути искуственного повышения рейтинга внешнего сайта в поисковых машинах.

На рис. 4 показано использование микроформата XFN. Применение этого простого микроформата в блогах позволило создать такие инструменты, как *rubhub.com*, где реализована навигация по социальным связям между различными веб-ресурсами.

4 RDFa

Рассмотрим ещё одну технологию для обогащения гипертекста семантической информацией – RDFa [4]. RDFa – это метод семантической разметки с помощью нескольких дополнительных XHTML-атрибутов, значения которых выбираются из некоторого набора словарей, который может быть расширен. Как и рассмотренные выше методы, RDFa направлен на то, чтобы превратить информацию, предназначенную для прочтения человеком, в машинно-читаемую без дублирования данных, а лишь за счет дополнительной разметки.

RDFa более универсален и имеет несколько существенных отличительных особенностей по сравнению с микроформатами, хотя они и предназначены для выполнения той же задачи. Далее перечислены основные отличия от микроформатов:

- RDFa чаще использует новоопределенные артибуты XHTML и совсем не использует атрибут class;
- значения атрибутов не специфицируются централизованным стандартом, как в случае микроформатов, а берутся из различных словарей в сети, например, FOAF или DC;
- поскольку в разных словарях один и тот же термин может использоваться для обозначения различных сущностей, в RDFa для каждого словаря определяется своё пространство имен; в то же время микроформаты используют глобальное пространство имен для определяемых значений атрибутов;
- RDFa может расширить любой пользователь, для этого достаточно создать свой словарь и описать семантику вводимых значений для третьих лиц; напротив, создание нового микроформата осуществляется централизованно и контролируется одним сообществом;
- модель данных каждого микроформата жестко прописана в его спецификации и отличается для разных микроформатов; модель данных RDFa универсальна и базируется на модели данных RDF, то

```
<DIV CLASS="VEVENT">
	<a href="http://isofts.kiev.ua" class="url">
		<span class="summary">Конференция УкрПрог</span>
	</a> проводится
	<abbr title="2010-02-17" class="dtstart">17</abbr> -
	<abbr title="2010-02-21" class="dtend">21</abbr> февраля 2010 г.
	в <span class="location">Институте программных систем</span>
	<div class="description">Конференция посвящена проблемам программирования</div>
</DIV>
```

Рис. 1. Пример использования микроформата hCalendar

```
<div class="vcard">
	<span class="fn n">
		<span class="family-name">Кудим</span>
		<span class="given-name">Кузьма</span>
		<span class="additional-name">Алексеевич</span>
	</span>
	<div class="org">ИПС</div>
	<a class="email" href="mailto:kuzma@isofts.kiev.ua">kuzma@isofts.kiev.ua</a>
	<div class="adr">
		<span class="locality">Киев</span>,
		<span class="country-name">Украина</span>
	</div>
	<div class="tel">067-5889125</div>
</div>
```

Рис. 2. Пример использования микроформата hCard

```
<a href="http://myspam.ua" rel="nofollow">Посмотри, клёвая вещь!</a>
```

Рис. 3. Пример использования микроформата rel-nofollow

```
<a href="http://my.blog.ua" rel="me">Мой сайт</a>
<a href="http://some.blog.ua" rel="friend colleague neighbor">Платон</a> мне друг, коллега и сосед.
```

Рис. 4. Пример использования микроформата XFN

есть на представлении структуры данных в виде графа с помеченными ребрами, где узлы – это данные, а ребра, определяемые атрибутами, – это связи между данными.

Обобщая перечисленные отличия, можно сказать, что микроформаты покрывают только наиболее очевидные и часто используемые структуры в веб-документах, стараясь максимально упростить разметку этих структур. Идеологи микроформатов утверждают, что этого должно быть достаточно в подавляющем большинстве случаев. Совсем иной подход RDFa, который базируется не на ad hoc решениях, а на универсальной формальной модели, что несколько затрудняет его использование, но зато обеспечивает охват всевозможных структур данных.

Из примера на рис. 5 можно видеть, что синтаксис RDFa достаточно прост, от аналогичного примера для микроформата hCard (рис. 2) он отличается только повсеместным указанием префикса в значениях соответствующих атрибутов и именами этих атрибутов.

5 Встраиваемый RDF

Существует компромиссный вариант между жесткой структурой микроформатов, не использующих новых атрибутов, и гибкостью RDFa, которая достигается введением дополнительных атрибутов. Подход этот называется RDF, встраиваемый в XHTML (embeddable RDF, eRDF) [5].

По аналогии с микроформатами, eRDF не использует новых атрибутов XHTML. Для семантической разметки используются атрибуты class, id, rel, rev. Специальный синтаксис позволяет обрабатывать значения этих атрибутов и преобразовывать в RDF-тройки. Для иллюстрации возможностей разметки с его помощью ограничимся примером на рис. 6.

eRDF реализует только ту часть RDF, которую можно выразить ограниченными средствами, без введения новых атрибутов, но, тем не менее, этот метод более гибок, чем микроформаты.

6 XML

С одной стороны, как микроформаты, так и RDFa подвержены критике в той части, где они переносят структуру документа в атрибуты, тем самым скрывая её: микроформаты – из идейных соображений, чтобы не вносить никаких изменений в XHTML, а RDFa – из-за универсальности, чтобы можно было произвести разметку любой сложности с использованием

```
<div class="vcard" xmlns:v="http://www.w3.org/2001/vcard-rdf/3.0#">
    <span property="v:Family">Кудим</span>
    <span property="v:N">Кузьма</span>
    <span property="v:Given">Алексеевич</span>
    <div class="v:ORG">ИПС</div>
    <a rel="v:EMAIL" href="mailto:kuzma@isofts.kiev.ua">kuzma@isofts.kiev.ua</a>
    <div role="v:ADR">
        <span property="v:Locality">Киев</span>,
        <span property="v:Country">Украина</span>
    </div>
    <div role="v:TEL">
        <span property="v:Value">067-5889125</span>
    </div>
</div>
```

Рис. 5. Пример использования RDFa

```
<div id="kuzma">
       <span class="vcard-Family">Кудим</span>
       <span class="vcard-N">Кузьма</span>
       <span class="vcard-Given">Алексеевич</span>
       <div class="vcard-ORG">ИПС</div>
       <a class="vcard-EMAIL" href="mailto:kuzma@isofts.kiev.ua">kuzma@isofts.kiev.ua</a>
       <div class="vcard-ADR">
              <span class="vcard-Locality">Киев</span>,
              <span class="vcard-Country">Украина</span>
       </div>
       <div class="vcard-TEL">067-5889125</div>
</div>
```

Рис. 6. Пример использования eRDF

```
<vcard>
    <family>Кудим</family>
    <name>Кузьма</name>
    <given>Алексеевич</given>
    <org>ИПС</org>
  <email><a href="mailto:kuzma@isofts.kiev.ua">kuzma@isofts.kiev.ua</a></email>
    <adr>
        <locality>Киев</locality>,
        <country>Украина</country>
    </adr>
    <tel type="mobile">067-5889125</tel>
</vcard>
```

Рис. 7. Пример использования XML

любых словарей, опираясь на единый синтаксис. С другой стороны, стандарт XHMTL, как вообще принято для XML, отражает структуру данных непосредственно на уровне элементов разметки.

Учитывая, что XHTML является расширяемым модульным языком [5], можно реализовать разнообразные структуры внутри XHTML-разметки более элегантно. Рис. 7 иллюстрирует, насколько яснее становится разметка при наличии необходимых элементов. Здесь каждый элемент уже сам несёт смысловую нагрузку, заместив более общие элементы DIV и SPAN. Поскольку клиентским приложениям нововведенные элементы не известны, то для правильного отображения необходимо связать их с таблицей стилей.

Данный подход не является распространенным. В целом он близок к микроформатам: семантика элементов должна быть где-то централизованно описана, для каждого случая требуется вводить отличную структуру разметки.

Следует также отметить, что любой такой дополненный XHTML можно привести с помощью XSL-преобразований [6] к соответствующему микроформату либо RDFa.

7 Заключение

В связи с растущим интересом веб-сообщества к представлению информации не только для понимания человеком, но и для повышения эффективности ее автоматической обработки, в последнее десятилетие стали активно развиваться технологии по обогащению гипертекста дополнительной структу-

рой и семантическими метаописаниями. Два наиболее распространенных и стандартизованных на сегодняшний день метода семантической разметки XHTML – микроформаты и RDFa. Микроформаты создавались с прицелом на простоту, минималистичность дополнительной разметки и призваны обеспечить решение задачи семантической разметки в наиболее очевидных и распространенных случаях. Принцип RDFa – универсальность, возможность обеспечить всю полноту семантического представления в рамках гипертекста, что привело к некоторому расширению языка XHTML. Оба подхода можно упрекнуть в перенесении структуры документа в атрибуты, в то время как идеология XML, на котором основан XHTML, предполагает отражение структуры явно на уровне элементов разметки. В этой связи рассматривается еще одна возможность – введение в XHTML необходимых дополнительных элементов разметки.

Литература

[1] W3C Semantic Web Activity. – http://www.w3.org/2001/sw/.

[2] XHTML 1.1 – Module-based XHTML. – http://www.w3.org/TR/xhtml11/.

[3] Microformats. – http://microformats.org/.

[4] RDFa primer. Bridging the human and data Webs. – http://www.w3.org/TR/xhtml-rdfa-primer/.

[5] XHTML Modularization 1.1. – http://www.w3.org/TR/xhtml-modularization/.

[6] XSL Transformations (XSLT). – http://www.w3.org/TR/xslt.

Methods of web document semantic markup

K. Kudim

The paper is about methods of complementary semantic markup inside XHTML pages such as microformats, RDFa, eRDF and XML.