8 Информация о языке и направление текста

В этом разделе документа обсуждаются два важных вопроса интернационализации HTML: задание языка (атрибут `lang`) и направления (атрибут `dir`) текста в документе.

8.1 Задание языка содержимого: атрибут `lang`

Определения атрибутов

`lang` = *код языка* [CI]

Этот атрибут указывает основной язык значений атрибутов элементов и секстового содержимого. По умолчанию значение этого атрибута не установлено.

Информация о языке, указанная с помощью атрибута `lang`, может использоваться агентом пользователя для управления генерацией изображения различными способами. Некоторые ситуации, в которых указываемая автором информация о языке, может быть полезна:

- Помощь поисковым машинам
- Помощь синтезаторам речи
- Помощь агентам пользователей в выборе вариантов глифов для высококачественной типографии
- Помощь агенту пользователя в выборе набора кавычек
- Помощь агенту пользователя в вопросах переноса, лигатур и интервалов
- Помощь программа проверки грамматики и орфографии

Атрибут `lang` указывает код содержимого элемента и значений атрибутов; относится ли он к данному атрибуту, зависит от синтаксиса и семантики атрибута и от операции.

Атрибут `lang` предназначен для того, чтобы позволить агентам пользователей более осмысленно генерировать изображение на основе принятой культурной практики для данного языка. Это не подразумевает, что агенты пользователей должны генерировать символы, не являющиеся типичными для конкретного языка, менее осмысленным способом; агенты пользователей должны пытаться сгенерировать се символы, независимо от значения, указанного в атрибуте `lang`.

Например, если в русском тексте должен появиться символ греческого алфавита:

```
<P><Q lang="ru">"Эта супермощность была результатом &gamma;-радиации,</Q> объяснил он.</P>
```

агент пользователя (1) должен попытаться сгенерировать русский текст соответствующим образом (например, в соответствующих кавычках) и (2) попытаться сгенерировать символ γ, даже если это не русский символ.

Дополнительную информацию см. в разделе о неотображаемых символах.

8.1.1 Коды языков

Значением атрибута `lang` является код языка, идентифицирующий естественный разговорный язык, который устно, письменно или иным образом используется для передачи информации между людьми. Компьютерные языки явным образом исключены из кодов языков.

В документе [RFC1766] определены и описаны все коды языков, которые должны использоваться в документах на языке HTML.

Кратко говоря, коды языков состоят из первичного кода и ряда подкодов, который может быть пустым:

```
код-языка = первичный-код ( "-" подкод )*
```

Вот несколько примеров кодов языков:

- "en": английский
- "en-US": американская версия английского.

- "en-cockney": кокни (диалект английского).
- "i-navajo": навахо (язык американских индейцев).
- "x-klingon": Первичный код "x" обозначает экспериментальный код языка

Двухбуквенные первичные коды зарезервированы для сокращений языков по стандарту [ISO639]. Сюда входят коды fr (французский), de (немецкий), it (итальянский), nl (голландский), el (греческий), es (испанский), pt (португальский), ar (арабский), he (иврит), ru (русский), zh (китайский), ja (японский), hi (хинди), ur (урду) и sa (санскрит).
Любой двухбуквенный подкод считается кодом страны в стандарте [ISO3166].

8.1.2 Наследование кодов языков

Элемент наследует информацию о коде языка в следующем порядке старшинства (от высшего к низшему):

- Атрибут `lang`, установленный для самого элемента.
- Самый близкий родительский элемент, для которого установлено значение атрибута `lang` (то есть атрибут `lang` наследуется).
- Заголовок HTTP "Content-Language" (который может конфигурироваться на сервере). Например:

```
Content-Language: en-cockney
```

- Значения по умолчанию и настройки агента пользователя.

В этом примере первичным языком документа является французский ("fr"). Один абзац объявлен на испанском языке ("es"), после чего язык снова становится французским. В следующий абзац включена японская фраза ("ja"), после чего язык опять изменяется на французский.

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0//EN"
   "http://www.w3.org/TR/REC-html40/strict.dtd">
<HTML lang="fr">
<HEAD>
<TITLE>Un document multilingue</TITLE>
</HEAD>
<BODY>
...текст интерпретируется как французский...
<P lang="es">... текст интерпретируется как испанский...
<P>... текст опять интерпретируется как французский...
<P>...французский текст, в котором попадается
<EM lang="ja">фрагмент на японском</EM>, а здесь
опять начинается французский...
</BODY>
</HTML>
```

Примечание. *Ячейки таблицы могут наследовать значения атрибута `lang` не от родителя, а из первой ячейки объединения. Подробнее в разделе наследование выравнивания.*

8.1.3 Интерпретация кодов языков

В контексте HTML код языка должен интерпретироваться агентами пользователя как иерархия знаков, а не один знак. Если агент пользователя генерирует изображение в соответствии с информацией о языке (скажем, сравнивая языковые коды в таблицах стилей и значения атрибута `lang`), он всегда должен находить точное соответствие, но должен также принимать во внимание первичные коды. Таким образом, если значение атрибута `lang` "en-US" установлено для элемента `HTML`, агент пользователя должен сначала выбрать информацию о стиле, совпадающую с "en-US", а затем сгенерировать более общее значение "en".

Примечание. *Иерархия кодов языков не гарантирует понимания всех языков с общими префиксами людьми, бегло говорящими на одном или нескольких из этих языков. Она помогает пользователю запросить эту общность, когда для пользователя она является истинной.*

8.2 Указание направления текста и таблиц: атрибут

dir

Определения атрибутов

dir = LTR | RTL [CI]

Этот атрибут задает основное направление нейтрального в смысле направления текста (например, текста, который не наследует направленность, как определено в [UNICODE]) и направление таблиц. Возможные значения:

- LTR: Слева направо.
- RTL: Справа налево.

Кроме задания языка документа с помощью атрибута lang, авторы могут указать основное направление (слева направо или справа налево) частей текста, таблицы и т.д. Это делается с помощью атрибута dir.

Спецификация [UNICODE] назначает направление символам и определяет (сложный) алгоритм для определения соответствующего направления текста. Если документ не содержит отображаемых справа налево символов, агент пользователя не должен использовать двунаправленный алгоритм [UNICODE]. Если документ содержит такие символы, и если агент пользователя и отображает, он должен использовать двунаправленный алгоритм.

Хотя в Unicode определены специальные символы, отвечающие за направление текста, HTML предлагает конструкции разметки высшего уровня, выполняющие те же функции: атрибут dir (не спутайте с элементом DIR) и элемент BDO. Таким образом, чтобы привести цитату на иврите, проще написать

```
<Q lang="he" dir="rtl">...цитата на иврите...</Q>
```

чем с эквивалентными ссылками Unicode:

```
&#x202B;&#x05F4;...цитата на иврите...&#x05F4;&#x202C;
```

Агенты пользователей **не** должны использовать атрибут lang для определения направления текста.

Атрибут dir наследуется, и его можно переопределить. Подробнее см. в разделе о наследовании информации о направлении текста.

8.2.1 Введение в двунаправленный алгоритм

В следующем примере проиллюстрировано ожидаемое поведение двунаправленного алгоритма. В нем показаны английский текст слева направо и текст на иврите справа налево. Рассмотрите следующий текст:

```
  английский1 ИВРИТ2 английский3 ИВРИТ4 английский5 ИВРИТ6
```

Символы в этом примере (и во всех реплицированных примерах) хранятся в компьютере в том же виде, в каком они отображаются здесь: первый символ - "а", второй - "н", последний "6". Предположим, для содержащего этот абзац документа определен английский язык. Это означает, что основным направлением является направление слева направо. Корректное представление этой строки:

```
английский1 2ТИРВИ английский3 4ТИРВИ английский5 6ТИРВИ
             <-----              <-----              <-----
                H                  H                  H
-------------------------------------------------------->
                          E
```

Строки точек указывают структуру предложения: основным языком является английский, но встроены некоторые элементы на иврите. Для получения корректного представления не нужно никакой дополнительной разметки, поскольку фрагменты на иврите корректно обращаются агентами пользователя, применяющими двунаправленный алгоритм.

С другой стороны, если для документа определен язык иврит, основным будет направление справа налево. Корректное представление, таким образом, будет:

```
6ТИРВИ английский5 4ТИРВИ английский3 2ТИРВИ английский1
```

```
        ------->            ------->            ------->
           E                   E                   E
<-------------------------------------------------
                        H
```

В этом случае все предложение представляется справа налево, а фрагменты на английском языке обращаются двунаправленным алгоритмом.

8.2.2 Наследование информации о направлении текста

Для двунаправленного алгоритма Unicode необходимо основное направление текста для текстовых блоков. Чтобы указать основное направление элементов уровня блока, установите для этого элемента атрибута dir. Значением атрибута dir, устанавливаемым по умолчанию, является "ltr" (слева направо). Если атрибут dir установлен для элемента уровня блока, он действует на протяжении всего элемента и для всех вложенных элементов уровня блока. Установка атрибута dir для вложенного элемента имеет приоритет по отношению к наследуемому значению. Чтобы установить основное направление текста для всего документа, установите атрибут dir в элементе HTML.

Например:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0//EN"
   "http://www.w3.org/TR/REC-html40/strict.dtd">
<HTML dir="RTL">
<HEAD>
<TITLE>...заголовок справа налево...</TITLE>
</HEAD>
...текст справа налево...
<P dir="ltr">...текст слева направо...</P>
<P>...опять текст справа налево...</P>
</HTML>
```

Встроенные элементы, с другой стороны, не наследуют атрибут dir. Это означает, что встроенный элемент без атрибута dir **не** открывает дополнительного уровня внедрения в соответствии с двунаправленным алгоритмом. (Здесь элементом считается элемент уровня блока или встроенный элемент на основе представления по умолчанию. Помните, что элементы INS и DEL могут быть элементами уровня блока или встраиваемыми элементами в зависимости от контекста.)

8.2.3 Установка направления внедренного текста

Двунаправленный алгоритм [UNICODE] автоматически обращает последовательности внедренных символов в соответствии с наследуемым направлением (как показано в предыдущих примерах). Однако в общем в расчет принимается только один уровень внедрения. Для того чтобы изменения направления достигали дополнительных уровней, используйте атрибут dir во встроенном элементе.

Рассмотрите текст предыдущего примера:

```
английский1 ИВРИТ2 английский3 ИВРИТ4 английский5 ИВРИТ6
```

Предположим, основным языком для документа, содержащего этот абзац, является английский. В этом английском предложении содержится фрагмент на иврите, продолжающийся от ИВРИТ2 до ИВРИТ4, и в нем содержится англоязычный фрагмент (английский3). Таким образом, желаемое представление текста:

```
английский1 4ТИРВИ английский3 2ТИРВИ английский5 6ТИРВИ
                    ------->
                       A
             <-----------------------
                       И
-------------------------------------------------->
                       A
```

Для изменения направления текста двух внедренных фрагментов необходимо задать дополнительную информацию, что мы и делаем, явно разделяя второе внедрение. В этом

примере мы используем для разметки текста элемент SPAN и атрибут dir:

```
английский1 <SPAN dir="RTL">ИВРИТ2 английский3 ИВРИТ4</SPAN> английский5 ИВРИТ6
```

Авторы также могут использовать для изменения направления нескольких внедренных фрагментов символы Unicode. Для указания направления слева направо во внедряемом фрагменте окружите текст символами LEFT-TO-RIGHT EMBEDDING ("LRE", шестнадцатеричный код 202A) и POP DIRECTIONAL FORMATTING ("PDF", шестнадцатеричный код 202C). Для указания направления справа налево во внедряемом фрагменте окружите текст символами RIGHT-TO-LEFT EMBEDDING ("RTE", шестнадцатеричный код 202B) и PDF.

Использование разметки направленности HTML с символами Unicode. *Авторы и разработчики средств создания HTML-документов должны знать о возможных конфликтах, возникающих при использовании атрибута dir со встроенными элементами (включая BDO) одновременно с соответствующими символами форматирования [UNICODE]. Предпочтительнее использовать только один метод. Метод с использованием разметки гарантирует структурную целостность документа и устраняет некоторые проблемы с редактированием двунаправленного текста HTML в простых текстовых редакторах, но некоторое программное обеспечение может лучше использовать символы [UNICODE]. Если используются оба метода, следует хорошо позаботиться о правильном вложении разметки и символов, иначе результаты могут быть непредсказуемыми.*

8.2.4 Приоритет над двунаправленным алгоритмом: элемент BDO

```
<!ELEMENT BDO - - (%inline;)*          -- приоритет над I18N BiDi -->
<!ATTLIST BDO
  %coreattrs;                          -- id, class, style, title --
  lang        %LanguageCode; #IMPLIED  -- код языка --
  dir         (ltr|rtl)      #REQUIRED -- направление --
  >
```

Начальный тег: ***обязателен****, Конечный тег:* ***обязателен***

Определения атрибутов

dir = LTR | RTL [CI]

Этот обязательный атрибут указывает основное направление текстового содержимого элемента. Это направление имеет приоритет по отношению к наследуемому направлению символов, как определено в [UNICODE]. Возможные значения:

- LTR: Направление слева направо.
- RTL: Направление справа налево.

Атрибуты, определяемые в любом другом месте

- lang (информация о языке)

Двунаправленного алгоритма и атрибута dir обычно достаточно для управления изменением направления внедренного текста. Однако в некоторых ситуациях двунаправленный алгоритм может привести к некорректному представлению. Элемент BDO позволяет авторам отключать двунаправленный алгоритм для выбранных фрагментов текста. Рассмотрите документ с тем же текстовым фрагментом:

```
английский1 ИВРИТ2 английский3 ИВРИТ4 английский5 ИВРИТ6
```

но предположите, что этот текст уже представлен в нужном порядке. Одной причиной этого может быть то, что стандарт MIME ([RFC2045], [RFC1556]) благоприятствует визуальному порядку, то есть последовательности с направлением справа налево вставляются в байтовый поток с направлением справа налево. В электронной почте это может форматироваться, включая перевод строки, например:

```
английский1 2ТИРВИ английский3
```

```
4ТИРВИ английский5 6ТИРВИ
```

Это конфликтует с двунаправленным алгоритмом [UNICODE], поскольку этот алгоритм инвертирует 2ТИРВИ, 4ТИРВИи 6ТИРВИ во второй раз, так что слова на иврите отображаются слева направо, а не справа налево. Решением будет переопределить действие двунаправленного алгоритма, поместив выдержку Email в элемент PRE (для сохранения переводов строки) и каждую строку, для которой атрибут dir установлен в LTR, в элемент BDO:

```
<PRE>
<BDO dir="LTR">английский1 2ТИРВИ английский3</BDO>
<BDO dir="LTR">4ТИРВИ английский5 6ТИРВИ</BDO>
</PRE>
```

Двунаправленному алгоритму выдается команда "Я должен быть слева направо!", что приведет к нужному представлению:

```
английский1 2ТИРВИ английский3
4ТИРВИ английский5 6ТИРВИ
```

Элемент BDO следует использовать в сценариях, где необходим абсолютный контроль над последовательностью (например, многоязыковые номера частей). Атрибут dir для этого элемента является обязательным.

Авторы могут также использовать специальные символы Unicode для того, чтобы избежать использования двунаправленного алгоритма -- LEFT-TO-RIGHT OVERRIDE (202D) или RIGHT-TO-LEFT OVERRIDE (шестнадцатеричный код 202E). Символ POP DIRECTIONAL FORMATTING (шестнадцатеричный код 202C) заканчивает любую последовательность, используемую для обхода двунаправленного алгоритма.

Примечание. *Помните, что при использовании атрибута dir во встроенных элементах (включая BDO) одновременно с соответствующими символами форматирования [UNICODE], могут возникать конфликты.*

Двунаправленность и кодировка символов *В соответствии с [RFC1555] и [RFC1556] существуют специальные соглашения относительно использования значений параметра "charset" для указания обработки двунаправленности в почте MIME, в частности для отличия визуальной, явной и неявной направленности. Значение параметра "ISO-8859-8" (для иврита) обозначает визуальную кодировку, "ISO-8859-8-i" обозначает неявную двунаправленность, а "ISO-8859-8-e" обозначает явную направленность.*

Поскольку HTML использует двунаправленный алгоритм Unicode, соответствующие документы, использующие кодировку ISO 8859-8, должны помечаться как "ISO-8859-8-i". Явное управление направлением в HTML также возможно, но его нельзя выразить в ISO 8859-8, поскольку не следует использовать "ISO-8859-8-e". Значение "ISO-8859-8" подразумевает, что документ отформатирован визуально, и некоторая разметка будет использоваться неправильно (например, TABLE с выравниванием по правому краю без разбивки строк), чтобы гарантировать правильное отображение для более старых агентов пользователя, не поддерживающих двунаправленность. Такие документы не удовлетворяют настоящей спецификации. При необходимости их можно изменить (и одновременно они будут корректно отображаться в старых версиях агентов пользователей), добавив, где нужно, разметку BDO. Вопреки сказанному в [RFC1555] и [RFC1556], кодировка ISO-8859-6 (арабская) ***не*** *представляет визуального порядка.*

8.2.5 Ссылки на символы для управления направлением и объединением

Поскольку иногда возникает двусмысленность относительно некоторых символов (например, символов пунктуации), спецификация [UNICODE] включает символы для правильного определения назначения. Спецификация Unicode также включает некоторые символы для управления объединением при необходимости (например, в ситуациях с арабскими символами). HTML 4.0 включает для этих символов ссылки на символы.

Следующее DTD определяет представление некоторых объектов направления:

```
<!ENTITY zwnj CDATA "&#8204;"--=нулевая ширина без объединения -->
<!ENTITY zwj  CDATA "&#8205;"--=объединитель нулевой ширины-->
<!ENTITY lrm  CDATA "&#8206;"--=метка слева направо-->
<!ENTITY rlm  CDATA "&#8207;"--=метка справа налево-->
```

Объект `zwnj` используется для блокировки объединения в тех контекстах, где объединение произойдет, но оно происходить не должно. Объект `zwj` имеет обратное действие; он производит объединение в случае, когда оно не предполагается, но должно произойти. Например, арабская буква "НЕН" используется для сокращения "Hijri", названия исламской системы летоисчисления. Поскольку отдельный иероглиф "НЕН" в арабской письменности выглядит как цифра пять, для того, чтобы не путать букву "НЕН" с последней цифрой пять в годе, используется исходная форма буквы "НЕН". Однако, нет последующего контекста (например, буквы для объединения), с которым можно объединить "НЕН". Символ `zwj` предоставляет такой контекст.

Точно так же в персидских текстах буква может иногда объединяться с последующей буквой, в то время как в рукописном тексте этого быть не должно. Символ `zwnj` используется для блокировки объединения в таких случаях.

Символы порядка, `lrm` и `rlm`, используются для определения направления нейтральных по отношению к направлению символов. Например, если двойные кавычки ставятся между арабской (справа налево) и латинской (слева направо) буквами, направление кавычек неясно (относятся ли они к арабскому или к латинскому тексту?). Символы `lrm` и `rlm` имеют свойство направления, но не имеют свойств ширины и разделения слов/строк. Подробнее см. [UNICODE].

Отражение глифов символов. *Вообще двунаправленный алгоритм не отражает глифы символов и не влияет на них. Исключением являются такие символы как скобки (см. [UNICODE], таблица 4-7). Если отражение желательно, например, для египетских иероглифов, греческих знаков или специальных эффектов дизайна, можно сделать это с помощью стилей.*

8.2.6 Таблицы стилей и двунаправленность

Вообще использование таблиц стилей для изменения визуального представления элемента с уровня блока до встроенного и наоборот используется в прямом направлении. Однако, поскольку двунаправленный алгоритм использует различия встроенных элементов/элементов уровня блока, во время преобразования нужно быть внимательным.

Если встроенный элемент, не имеющий атрибута `dir,` преобразуется в стиль элемента уровня блока с помощью таблицы стилей, для определения основного направления блока он наследует атрибут `dir` от ближайшего родительского элемента блока.

Если элемент блока, не имеющий атрибута а `dir,` преобразуется к стилю встроенного элемента с помощью таблицы стилей, результирующее представление должно быть эквивалентным, в терминах двунаправленного форматирования, форматированию, получаемому путем явного добавления атрибута `dir` (которому назначено унаследованное значение) преобразованному элементу.