2E-5

# HTML 及び XML における Bidi 問題の解決手法

堀江 史郎　和田 雄策　伊藤 一成　Martin J. DÜRST
青山学院大学理工学部

## 1　はじめに

現在世界には約 6500 種類の言語があり，日本語や英語などの左から右に記述する文字（以下左からとする）を使う言語と，アラビア語やヘブライ語などの右から左に記述する文字（以下右からとする）を使う言語がある [1]．左からの文字と右からの文字が同一の文に混在した際の，一般文書の表示方法については，Unicode Bidi アルゴリズムにて規定されている [2]．しかし，C 言語や Java，HTML，XML などのソースの場合は問題が発生する．山括弧（<, >）や引用符（', "）などの記号が原因となりソースの構造が正しく表示されない（図 1 参照）．このためソースの編集がほぼ不可能となる．これを Bidi 問題という．

本稿では，アラビア語やヘブライ語を使用している多くの人を悩ます，HTML 及び XML における Bidi 問題を解決する．HTML 及び XML はプログラム言語に比べてヘブライ語やアラビア語の使用頻度が高い．これを解決すればプログラム言語にも応用できると考えられる．

Bidi 問題の解決を HTML のマークアップによって試みる．HTML を使うとソースの表示をブラウザでシミュレートでき，CGI を公開して多くの人が試用できる．本 CGI は改良に改良を重ね，以前 [3] に比べいくつか機能を追加している．

## 2　Unicode Bidi アルゴリズム

Unicode Bidi アルゴリズムは Bidi 問題の分析と解決に必要な，文字の方向性を制御する要素を提供しているため，本章で概要を説明する．

アルゴリズムでは，まず文字の種類によって文字の方向性を決める．それをもとに，基本方向（文全体の方向性）と異なる方向性をもつ文字列がある場合は，その文字列を左右反転させて表示する（図 2 参照）．

文字の方向性には，強い左からの文字（平仮名やラテン文字など）と 強い右からの文字（アラビア文字やヘブライ文字など），中性文字（各種記号や空白）がある．中性文字は決まった方向性をもたず，基本方向を継承する．ただし，両側に基本方向と異なる方向性をもつ強い文字がきた場合は，その方向性が適用される．

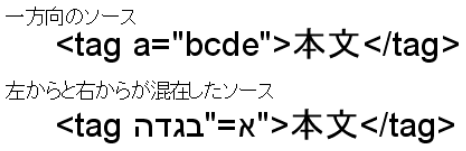


図 1: 一方向と混在した XML のソース

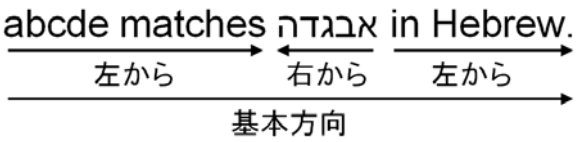


図 2: Unicode Bidi アルゴリズムの適用例

文の表示の一部を直すには LRM と RLM の制御文字を使用する．それぞれ強い左からの文字，強い右からの文字として働く．文字幅が 0 であり，一般には表示されない．

文全体と方向性が異なる文を埋め込むなど，LRM と RLM で不十分な場合はエンベッディングを使用する．エンベッディングを使うと，一定の範囲の文字列に対して左から，または右からの基本方向を付与できる．範囲指定の開始はそれぞれ制御文字 LRO，RLO で表し，終点は制御文字 PDF で表す．

文字の方向性を完全に無視して文字列を表示させたい場合はオーバーライドを使用する．オーバーライドを使うと，一定範囲の文字の方向性を強制的に設定できる．範囲指定の開始はそれぞれ制御文字 LRO，RLO で表し，終点は制御文字 PDF で表す．

一般文書では各制御文字を利用するが，HTML 仕様書の 8 章 2 節によると制御文字に替わりマークアップが使用される [5]．マークアップは制御文字と違いソースで表示されるので，使用箇所が明瞭であるなどの利点がある．

**A method for solving the bidirectional problem for HTML and XML source**
Shiro HORIE, Yusaku WADA, Kazunari ITO and Martin J. DÜRST
College of Science and Engineering, Aoyama Gakuin University
5–10–1 Fuchinobe, Sagamihara, Kanagawa 229-8558, Japan

<tag א="בגדה">本文</tag>

図 3: 図 1 の解決例

## 3 Bidi 問題の分析と解決手法

HTML 及び XML における Bidi 問題を探り，3 つの段階の問題を発見した．はじめに，最大の問題となっている “中性文字の方向”，次に，より使いやすくするための “ソースの基本方向”，最後に HTML 及び XHTML のみに起こる “HTML ソースでの方向指定の反映” である．

### 3.1 中性文字の方向

HTML 及び XML ではソースの構造を記述するために記号を多用するため，記号の割合が一般文書に比べて多くなる．この場合 Unicode Bidi アルゴリズムを適用しただけではソースの表示が崩れてしまう．これは，タグを構成する記号が周りの強い文字から影響を受けて，基本方向と異なる方向性を持ったものができるためである．

これを解決するために，制御文字 LRM を使用する．タグを構成している記号の両側に LRM を挿入することで，中性文字に方向性を与え，方向を統一させる．これにより，中性文字の方向の問題は解決した（図 3 参照）．HTML では LRM のマークアップである `&lrm;` を用いる．例えば，対象のソース（以下表のソースとする）が `<tag>` である場合，表のソースを Web で表示するための一段階深いソース（以下裏のソースとする）は `&lt;tag&gt;` になる．それに `&lrm;` を追加すると，`&lrm;&lt;&lrm;tag&lrm;&gt;&lrm;` となる．それをブラウザで閲覧することで，中性文字の方向を制御するシミュレーションを行う．

### 3.2 ソースの基本方向

これまでソースの基本方向を左からと想定していた．しかし，ソースの内容によっては，基本方向を右からにしたほうが読みやすくなる可能性がある．例えば，右からの文字を多く含む文章や，タグに右からの文字を多用したソースなどである．また，基本方向を統一せず，タグなどソースの一部を逆の方向性にするほうが読みやすいなども考えられる．その上，個人の好み，習慣などによって見やすい表示が異なる可能性もある．なお，ソースの基本方向の制御にはエンベッディングを使う．

以上のように，ソースの方向性を一定にするのは難しく，方向性を内容や利用者の好みに応じて変更できることが望ましいと考えられる．ソースを読みやすくするために必要と考えられる設定項目を調査する．

### 3.3 HTML ソースでの方向指定の反映

HTML 及び XHTML にはソースの方向性を指定するためのマークアップが用意されている．マークアップは Web ページとして表示した際に効果が現れるが，ソースの編集時には効果が現れない．そのため表示の順番が異なり，混乱を引き起こす可能性がある．例えば，文字列 “abc” に HTML で右からのオーバーライドを使う．このとき表のソースで `<bdo dir="rtl">abc</bdo>` と記述され，Web ページで “cba” と表示される．

この問題を解決するためにソースの分析を行う．その結果，表のソースに方向性が使われた際に，その方向性に対応させ裏のソースで方向性の制御を行う．この裏のソースをブラウジングして表のソースを表示することで，表のソースの表示を Web ページの表示に合致させることができる．

## 4 おわりに

本稿では，Bidi 問題を解決するため，3 つの問題について分析し，方向性の異なる文字が混在した際の表示を改善する手法を提案した．今後は，この手法を既存のエディタに実装することで，Bidi 問題を解決したエディタの実現を目指す．

この手法を体験できる Web ページを公開している [4]．ぜひ試用していただきたい．

## 参考文献

[1] 三上善貴: 世界の文字と文字符号（前編）, 情報処理学会学会誌 Vol. 46 No. 8 pp. 919-924 (2005).

[2] Mark Davis: Unicode Standard Annex #9, The Bidirectional Algorithm.
http://www.unicode.org/reports/tr9/

[3] Martin J. Dürst, Shiro Horie, Yusaku Wada: Exploring Better Source Editing for Bidirectional XHTML and XML, September 2005.

[4] Martin J. Dürst, Shiro Horie, Yusaku Wada: Bidi Source Editing Test CGI
http://www.sw.it.aoyama.ac.jp/2005/bidi-source

[5] Dave Raggett, Arnaud Le Hors, Ian Jacobs: HTML 4.01 Specification
http://www.w3.org/TR/html401/