
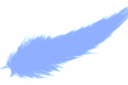


Implementing Better Source Editing for Bidirectional HTML and XML in the Text Editor Emacs


35th Internationalization and Unicode Conference
October 18, 2011
Shunsuke Oshima
Martin J. Dürst
Aoyama Gakuin University, Japan




Location of Talk/Slides/Software/Demos

<http://www.sw.it.aoyama.ac.jp/2011/pub/IUC35/>

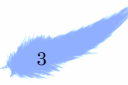
October 18, 2011 Internationalization and Unicode Conference 35






Contact

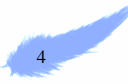
- Martin J. Dürst:
duerst@sw.it.aoyama.ac.jp
- Shunsuke Oshima:
shunshun9460 at gmail dot com


October 18, 2011 Internationalization and Unicode Conference 35  3




Overview


- Motivation
- Background Knowledge
- Analysis
- Implementation
- Conclusion


October 18, 2011 Internationalization and Unicode Conference 35  4



Motivation


- Ideal display:
`<p>Hello, ירושלים</p>`
- Actual display:
`<p>Hello, </p>`
- Unworkable!

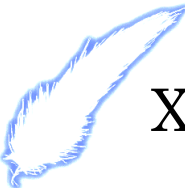
October 18, 2011 Internationalization and Unicode Conference 35  5



Background


- XML
- Bidirectionality
- Emacs


October 18, 2011 Internationalization and Unicode Conference 35  6



XML


- Fundamental Web Technology
 - XHTML for Web pages
 - SVG and X3D for 2D and 3D graphics
 - Many more
- Uses tags to mark up document/data structure

October 18, 2011 Internationalization and Unicode Conference 35  7



Example of XML

```
<?xml version="1.0" encoding="UTF-8"?>
<people>
  <person gender="male">
    <name>Charlie</name>
    <hobby>playing guitar</hobby>
  </person>
  <person gender="female">
    <name>Susan</name>
    <hobby>reading books</hobby>
  </person>
</people>
```

October 18, 2011 Internationalization and Unicode Conference 35  8



Bidirectionality

- Bidirectional characters
 - Left to Right ŪřƳIIIकधनेःC∞|ほ島
 - Right to Left |٩' البيان
- Unicode Bidirectional Algorithm
 - Display rules for bidirectional characters
 - For running text
 - Newspaper articles
 - Letters

October 18, 2011

Internationalization and Unicode Conference 35

9



Different kinds of text

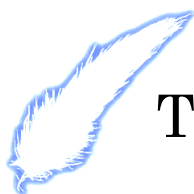
- Running text
 - Letters, newspaper articles,...
 - Bidi algorithm mostly adequate
 - Control characters can be inserted
- Structured text
 - XML, TeX, programming languages,...
 - Bidi algorithm highly inadequate
 - Control characters are invalid

October 18, 2011

Internationalization and Unicode Conference 35

10





Three Steps to Emacs Fun

- What is Emacs?
- What, Emacs?
- Emacs, of course!

October 18, 2011

Internationalization and Unicode Conference 35

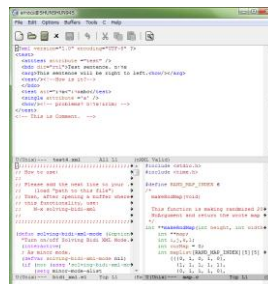
11



What is Emacs?



- Plain text editor
- Widely used
 - Long history
 - Uncountable functions
 - Amazing extensibility and customizability
- Extend by programming language Emacs Lisp



October 18, 2011

Internationalization and Unicode Conference 35

12





What, Emacs?

- Too complicated, antiquated, boring,...?
- No, not really:
 - Menus, dialogues, syntax highlighting
 - Available for many OSes (incl. Windows)
 - Even includes some games

October 18, 2011

Internationalization and Unicode Conference 35

13



Emacs and Internationalization

- Started with nemacs and mule
- Currently (23.3): Internal encoding based on UTF-8
- New in 24.0 (alpha/preview):
bidi reordering

October 18, 2011

Internationalization and Unicode Conference 35

14



(for reference: installing Emacs 24.0 on Windows)

- Download and unzip latest prerelease from <http://alpha.gnu.org/gnu/emacs/windows/>
- Run bin/addpm.exe for Start Menu entry
- Run Emacs from Start Menu
- Use Options → Multilingual Environment → Show Multi-lingual
- Explore and have fun

October 18, 2011

Internationalization and Unicode Conference 35

15




Analysis

October 18, 2011

Internationalization and Unicode Conference 35

16



Why the problem happens

- Syntactically significant characters are weak or neutral
- Between RTL characters, they become part of an RTL run.

```
<p>Hello, ירושלים</p>
```

```
<p>Hello, שלום</p>
```

October 18, 2011
Internationalization and Unicode Conference 35
17



Goal: Fix it!

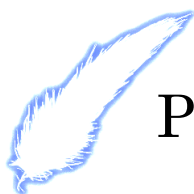
```
<p>Hello, שלום</p>
```



```
ירושלים'>שלום
```

```
<p>Hello, ירושלים</p>
```

October 18, 2011
Internationalization and Unicode Conference 35
18



Previous Research

- 2005: Web-based simulator
Problem: not interactive
- 2008: JavaScript implementation
Problems: brittle,
difficulties with local files

October 18, 2011

Internationalization and Unicode Conference 35

19



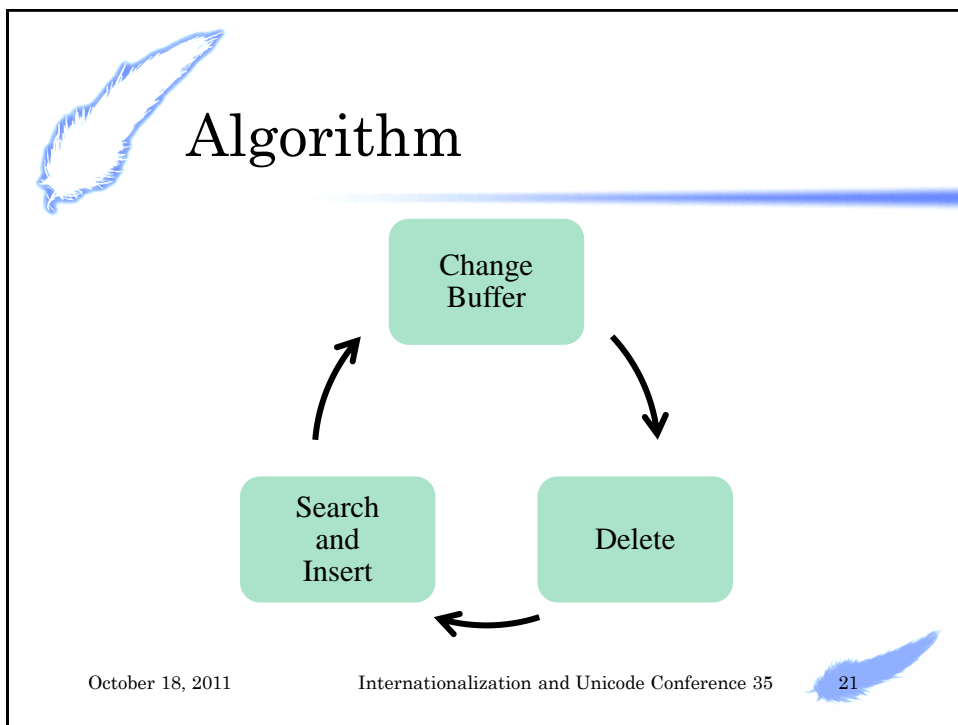
Implementation overview

- To fix display, temporarily insert implicit direction marks (LRM or RLM)
- Important to identify these marks and removed them e.g. before saving

October 18, 2011

Internationalization and Unicode Conference 35

20



- # Algorithm
1. Start whenever buffer is changed
 2. Delete all inserted implicit directional marks in current buffer
 3. Search for places where marks need to be inserted
 4. Insert the marks
- October 18, 2011 Internationalization and Unicode Conference 35 22



Character Insertion Details

- Extensibility
 - With small fix, able to fit most situation
- Not easy to implement
 - From top to bottom inserting is difficult
 - Reversed inserting is not difficult

October 18, 2011

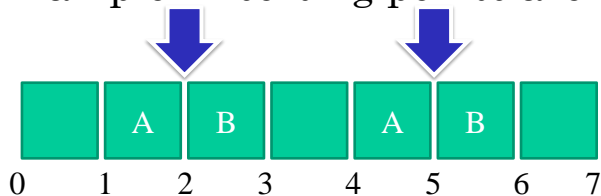
Internationalization and Unicode Conference 35

23



Way to insert

Example: Inserting points are 2 and 5




First, insert at 5

October 18, 2011

Internationalization and Unicode Conference 35

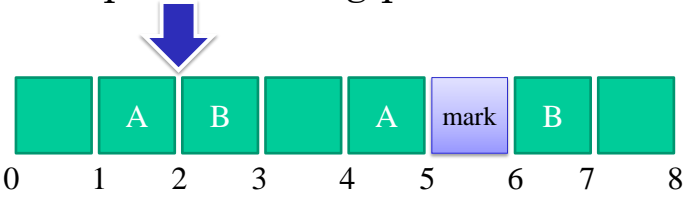
24






Way to insert

Example: Inserting points are 2 and 5



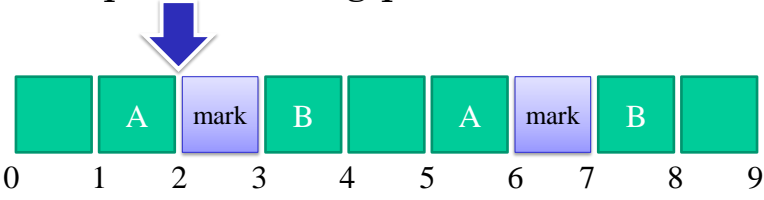
Then, insert to 2.

October 18, 2011 Internationalization and Unicode Conference 35 25




Way to insert

Example: Inserting point is 2 and 5.



If 2 is inserted first, the second A and B move


October 18, 2011 Internationalization and Unicode Conference 35 26



Syntactic Constructs in XML

- ❑ Start Tags
- ❑ End Tags
- ❑ Empty Tags
- ❑ Attributes
- ❑ Comments
- ❑ Processing Instructions
- ❑ CDATA Sections
- ❑ Document Type Definition

October 18, 2011 Internationalization and Unicode Conference 35 27

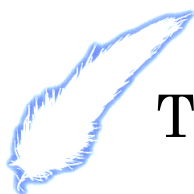


Comments

No effect to display

```
[LRM]<!-- [LRM] Comment [LRM]-->[LRM]
```

October 18, 2011 Internationalization and Unicode Conference 35 28



Tags

- Start Tags
- End Tags
- Empty Tags

```
[LRM]<[LRM] Start [LRM]>[LRM]  
[LRM]</[LRM] End [LRM]>[LRM]  
[LRM]<[LRM] Empty-element [LRM]/>[LRM]
```

October 18, 2011

Internationalization and Unicode Conference 35

29



Attributes

Add information to Elements

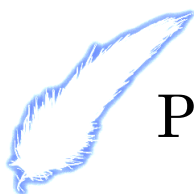
Two ways to quote values

```
name[LRM]="[LRM] value [LRM]"[LRM]  
name[LRM]='[LRM] value [LRM]'[LRM]
```

October 18, 2011

Internationalization and Unicode Conference 35

30



Processing Instructions

Add information to XML

- Version
- Encoding
- Style Sheet

```
[LRM]<?[LRM] PI [LRM]?>[LRM]
```

October 18, 2011

Internationalization and Unicode Conference 35

31



CDATA Sections

Literal data, not interpreted as markup

```
[LRM]<![CDATA[LRM] CDATA [LRM]]>[LRM]
```

October 18, 2011

Internationalization and Unicode Conference 35

32



Document Type Definition

Declare structure of XML

```
[LRM]<!DOCTYPE[LRM] DTD [LRM]>[LRM]
```

October 18, 2011

Internationalization and Unicode Conference 35

33



HTML dir attribute

- In the bdo element, dir means override
`<bdo dir="rtl">This text is an example.</bdo>`

.elpmaxe na si txet sihT




- We changed the display to
`<bdo dir="rtl">.elpmaxe na si txet sihT</bdo>`
- In other elements, dir means embedding

October 18, 2011

Internationalization and Unicode Conference 35

34



Way to solve dir attribute


search for dir attribute

```
<bdo dir="rtl">  
  <div>Reversed Display<br/>  
    <span>foo</span>  
  </div>  
</bdo>
```

Source

Stack

October 18, 2011 Internationalization and Unicode Conference 35 35



Way to solve dir attribute

Push bdo on stack and insert first RLO


```
<bdo dir="rtl">RLO  
  <div>Reversed Display<br/>  
    <span>foo</span>  
  </div>  
</bdo>
```

Source

Stack

bdo

October 18, 2011 Internationalization and Unicode Conference 35 36



Way to solve dir attribute

Close and reopen RLO at line break

```

<bdo dir="rtl">RLOPDF
RLO  <div>Reversed Display<br/>
      <span>foo</span>
    </div>
</bdo>



```

Source

Stack

bdo

October 18, 2011
Internationalization and Unicode Conference 35
37

Way to solve dir attribute

Push div on stack and insert marks

```

<bdo dir="rtl">RLOPDF
RLO  PDF<div>RLOReversed Display<br/>
      <span>foo</span>
    </div>
</bdo>


```

Source

Stack

bdo
div

October 18, 2011
Internationalization and Unicode Conference 35
38





Way to solve dir attribute

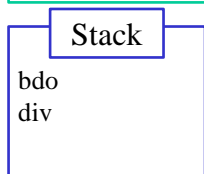
Insert marks around br but do not push

```

<bdo dir="rtl">RLOPDF
RLO  PDF<div>RLOReversed DisplayPDF<br/>RLO
      <span>foo</span>
      </div>
</bdo>

```

Source



October 18, 2011

Internationalization and Unicode Conference 35

39



Way to solve dir attribute

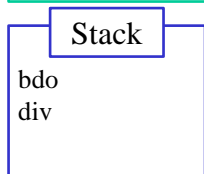
Close and reopen RLO at line break

```

<bdo dir="rtl">RLOPDF
RLO  PDF<div>RLOReversed DisplayPDF<br/>RLOPDF
RLO      <span>foo</span>
      </div>
</bdo>

```

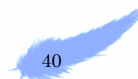
Source




October 18, 2011

Internationalization and Unicode Conference 35

40





Way to solve dir attribute

Push span on stack and insert marks

```

<bdo dir="rtl">RLOPDF
RLO  PDF<div>RLOReversed DisplayPDF<br/>RLOPDF
RLO      PDF<span>RLOfoo</span>
          </div>
</bdo>


```

Source

Stack

bdo
div
span

October 18, 2011 Internationalization and Unicode Conference 35 41



Way to solve dir attribute

Pop span from stack and insert marks

```

<bdo dir="rtl">RLOPDF
RLO  PDF<div>RLOReversed DisplayPDF<br/>RLOPDF
RLO      PDF<span>RLOfooPDF</span>RLO
          </div>
</bdo>


```

Source

Stack

bdo
div

October 18, 2011 Internationalization and Unicode Conference 35 42



Way to solve dir attribute

Close and reopen RLO at line break

```

<bdo dir="rtl">RLOPDF
RLO  PDF<div>RLOReversed DisplayPDF<br/>RLOPDF
RLO      PDF<span>RLOfooPDF</span>RLOPDF
RLO      </div>
</bdo>


```

Source

Stack

bdo
div

October 18, 2011 Internationalization and Unicode Conference 35 43



Way to solve dir attribute

Pop div from stack and insert marks

```

<bdo dir="rtl">RLOPDF
RLO  PDF<div>RLOReversed DisplayPDF<br/>RLOPDF
RLO      PDF<span>RLOfooPDF</span>RLOPDF
RLO  PDF</div>RLO
</bdo>

```

Source

Stack

bdo

October 18, 2011 Internationalization and Unicode Conference 35 44



Way to solve dir attribute

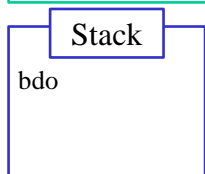
Close and reopen RLO at line break

```

<bdo dir="rtl">RLOPDF
RLO  PDF<div>RLOReversed DisplayPDF<br/>RLOPDF
RLO      PDF<span>RLOfooPDF</span>RLOPDF
RLO  PDF</div>RLOPDF
RLO</bdo>

```

Source



October 18, 2011

Internationalization and Unicode Conference 35

45



Way to solve dir attribute

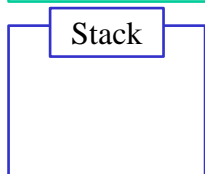
Pop bdo from stack and insert last PDF

```

<bdo dir="rtl">RLOPDF
RLO  PDF<div>RLOReversed DisplayPDF<br/>RLOPDF
RLO      PDF<span>RLOfooPDF</span>RLOPDF
RLO  PDF</div>RLOPDF
RLOPDF</bdo>

```

Source




October 18, 2011

Internationalization and Unicode Conference 35

46





Way to solve dir attribute

We are done!

```

<bdo dir="rtl">RLOPDF
RLO  PDF<div>RLOReversed DisplayPDF<br/>RLOPDF
RLO      PDF<span>RLOfooPDF</span>RLOPDF
RLO  PDF</div>RLOPDF
RLOPDF</bdo>

```

Stack Source

October 18, 2011 Internationalization and Unicode Conference 35 47



TeX/LaTeX

- TeX/LaTeX is a famous typesetting system
- In Bidi TeX, similar display problems:

```

\hello[San Francisco]
\hello[סלום ירושלים]

\hello{San Francisco}
\hello{סלום ירושלים}

```

October 18, 2011 Internationalization and Unicode Conference 35 48



Implementation details

- Implemented as major modes
- Major mode is a feature of Emacs
 - Used to customize editing for different file types
 - Syntax highlighting
 - File-type-specific functions
 - Users can add new major modes

October 18, 2011

Internationalization and Unicode Conference 35

49



Performance Improvement

- Limit insertion of control characters to area being displayed
- Performance does not degrade significantly for very long documents



text buffer

display

October 18, 2011

Internationalization and Unicode Conference 35

50



Demo

- From USB or CD
- Who wants to try?

- Please give the USBs back or hand them over to somebody else!
- You can keep the CD

October 18, 2011

Internationalization and Unicode Conference 35

51



Conclusion

- We solved the problem of bidirectional XML and HTML and for TeX/LaTeX for Emacs
- By using implicit directional marks

October 18, 2011

Internationalization and Unicode Conference 35

52



Future Work

- Debugging and integration
- Emacs-internal support
- More choices for users
- Other formats (CSS, programming languages,...)

October 18, 2011

Internationalization and Unicode Conference 35

53



Other editors

- Similar solution may be possible in other editors
- However,
 - Less extensible (no Emacs Lisp)
 - Maybe no bidi
- Hard, but sorely needed!

October 18, 2011

Internationalization and Unicode Conference 35

54