draft-ietf-iri-bidi-guidelines-03

IETF 85, Atlanta
IRI WG Meeting
2012-11-06

Martin J. Dürst, co-Editor

Overview

- Progress from -02 to -03
- Overview/Goals
- Problem 1: Separator Jumping
- Problem 2: External Confusion
- Problem 3: Visual Component Ordering
- Other open issues

Progress from -02 to -03
Non-ASCII Draft Versions

- New publication process to allow non-ASCII versions:
  - HTML: http://www.sw.it.aoyama.ac.jp/2012/pub/draft-ietf-iri-bidi-guidelines-03.html
  - PDF: http://www.ietf.org/id/draft-ietf-iri-bidi-guidelines-03.pdf
  - UTF-8 plaintext: http://www.sw.it.aoyama.ac.jp/2012/pub/draft-ietf-iri-bidi-guidelines-03.utf8.txt
  - ASCII only plaintext: http://www.ietf.org/id/draft-ietf-iri-bidi-guidelines-03.txt
- Please review a version with non-ASCII characters
- Please check for (bidi) rendering problems
- Please provide additional examples (more realistic)

Progress from -02 to -03
Closed Issues

- Issue #132: Allow non-spacing marks at end of components
- Issue #118: What term to use for the kind of text that the Unicode Bidi Algorithm was designed for
- Issue #116: logical order and 'read' order

Bidi(rectionality) Basics

- Arabic, Hebrew,… scripts read TFEL2THGIR
  (in examples, we use ESAC REPPU for right-to-left)
- Storage is in logical order (parsing and processing are easy)
- Display for general purpose text is specified by Unicode TR 9
  - Numbers are always left-to-right
  - Directionality of punctuation follows surrounding letters
  - In computer syntax, stuff gets thrown around
  - Demo/simulator available

Bidi IRI Goals

1. User-expected logical ⇔ display conversion
2. Uniform logical ⇔ display conversion
3. Low implementation cost (ideally just use TR 9)
4. Allow wide range of character combinations

Problem: Goals conflict, can’t have everything

Problem 1: Separator Jumping

- Logical Example: ONE1 . 2TWO
- Intended display? OWT2 . 1ENO
- Actual display! OWT1 . 2ENO
- Why: 1.2 is treated as a single number

Separator Jumping History

- Investigated in the context of IDNA 2008
- Extensive simulation by Harald Alvestrand (and Erik van der Poel)
- Led to RFC 5893: Rule for individual IDNA labels; if followed, very few surprises

Differences between IDNA and IRIs

- Separators:
  - IDNA: Dot only
  - IRIs: ‘:’, ‘/’, ‘.’, ‘@’, ‘?’, ‘#’, ‘=’, etc.
- Content:
  - IDNA: Internationalized LDH
  - IRIs: potentially anything (not only IDs but also payload)
- Control:
  - IDNA: separately per label
  - IRIs: often coordinated, sometimes separate or none

Separator Jumping for IRIs

- Investigation based on Harald's script
- Carried out by Shunsuke Oshima
  (Aoyama Gakuin University master student)
- Preliminary results:
  - '#' (class ET): more jumping than other separators
  - Minor differences between class CS (':', '.', '/',…) and class ON ('?', '@', '=', '&',…)
  - Constraints equivalent to those in IDNA should work reasonably well
- Detailed results to follow

Guidance to avoid Jumping

- If overall control is available: Check directly
- If overall control not available: Use heuristic equivalent to IDNA

Propose to add to document
(once investigation results are available)

Related: Issue #25: Adapt rules for bidi components to those in IDNA2008, Issue #28: allow numbers at end of bidi components?

Problem 2: External Confusion

- Imagine the following text (logical):
  LOOK AT THIS: new draft at http://ABC.COM
- May display as (in auto/RTL context):
  MOC.CBA//:new draft at http :SIHT TA KOOL
- Fix possible with bidi control characters (or markup) outside of IRI
- Users will fix, but maybe with bidi control characters inside IRI
- Issue #134: Check external influence on IRI display

Problem 3: Component Ordering

- Logical Example:
  def.com/ABC/XYZ.html
- Which display is intended?
  1. def.com/CBA/ZYX.html
  2. def.com/ZYX/CBA.html
  3. html.ZYX/CBA/com.def
  4. html.ZYX/CBA/def.com
- Good reasons for all of them !?

(Issue #121: Some users are requiring right-to-left label ordering)

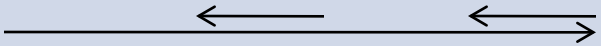
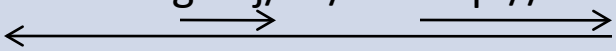
IRI Bidi Ordering Concepts

- Component: String between syntax characters
  - Domain name label
  - Path component
  - Query parameter name/value
  - …
- Component directionality:

  Each component clearly one way, to avoid ambiguities inside component and separator jumping
- Run: Same-directionality component sequence

Bidi IRI Ordering Alternatives

Logical: http://ab.CD/EF/gh?ij=KL#MN

Overall Directi-onality	Reordering by	Example	RFC 3987	Uni-code TR #9	Users	#
LTR →	run	http://ab.FE/DC/gh?ij=NM#LK 	okay	possible	☹	1
LTR →	component	http://ab.DC/FE/gh?ij=LK#NM 	bad	need ex-ception	☹	2
RTL ←	run	NM#LK=gh?ij/FE/DC.http://ab 	bad	possible	☹	3
RTL ←	component	NM#KL=ij?gh/FE/DC.ab//:http 	bad	need ex-ception	☺ ?	4

- Worst-case example
- Conflict between users (and user-oriented vendors) and security concerns

Exploring Solutions: Split ‘known location’ and running text

- Known locations:
  - Browser address bar,…
  - Side of a bus?
- (at least) two different ways to display
- Microsoft and Apple already started implementing
- Need for coordination:
  - Overall directionality?
  - Definition of components

Exploring Solutions: Determine overall directionality

- By context ⇒ two different orders
- By first component of domain name
  - Logical: ABC.def.ghi
  - Logical: def.ghi.ABC
  - Same display: def.ghi.CBA
- TLD component
  - Logical: abc.def.GHI
  - Logical: GHI.abc.def
  - Same display: abc.def.IHG
- Need advice to show TLD,… clearly

Exploring Solutions: Allow but remove bidi controls in IRIs

- Assumption: “optimal” bidi IRI display defined
- Fix running text display with bidi controls
- Remove bidi controls on resolution,…
- Cross-check display with bidi controls against “optimal” display, fail on non-match

Exploring Solutions: Guidance

- Try to keep things uniform:
  - All-RTL IRI
  - All-RTL (or LTR) domain name
  - All-RTL (or LTR) path
- Limit RTL components:
  - One domain label only
  - One path component only (e.g. Wikipedia)

Procedural Issue

- Issue #117: conformance requirements in bidi document - do they belong?
- Draft currently labeled as BCP
  - Right for advice on creation of bidi IRIs
  - Not appropriate for actual display rules

Acknowledgments

- Co-authors: Larry Masinter, Adil Allawi (عادل علاوي)
- Roozbeh Pournader, Aharon Lanin
- Shunsuke Oshima (大嶋 俊介)