Internationalized Resource Identifiers
(iri)
Internet-Draft
Intended status: BCP
Expires: April 24, 2013

M. Dürst
Aoyama Gakuin University
(青山学院大学)
L. Masinter
Adobe
A. Allawi (عادل علاوي)
Diwan Software Limited

October 21, 2012

TOC

Guidelines for Internationalized Resource Identifiers with Bi-directional Characters (Bidi IRIs) draft-ietf-iri-bidi-guidelines-03

Abstract

This specification gives guidelines for selection, use, and presentation of International Resource Identifiers (IRIs) which include characters with inherent right-to-left (rtl) writing direction.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as “work in progress.”

This Internet-Draft will expire on April 24, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF

Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

---

1. Introduction

TOC

1.1. Overview

TOC

Some UCS characters, such as those used in the Arabic and Hebrew scripts, have an inherent right-to-left (rtl) writing direction as opposed to characters, such as those in the Latin script, that have an inherent left-to-right (ltr) direction. IRIs containing rtl characters (called bidirectional IRIs or Bidi IRIs) require additional attention because of the non-trivial relation between their logical and visual ordering. The logical order represents the order in which characters are stored on computers and read by people. The visual order is the order in which the characters appear (or are expected to appear) on a computer display or printout.

Generally, alphabetic characters in scripts like Arabic and Hebrew are drawn rtl while numbers are drawn ltr. Symbols such as slash ('/') and period ('.') take their visual direction from the surrounding characters. A list of all ASCII symbols with their bidirectional character type and their function in URIs and IRIs is given in **Appendix A**.

Because of this complex interaction between the logical representation, the visual representation, and the syntax of a Bidi IRI, a balance is needed between various requirements. The main requirements are:

1. user-predictable conversion between visual and logical representation;
2.

the ability to include a wide range of characters in various parts of the IRI; and

3. minor or no changes or restrictions for implementations.

1.2. Availability

TOC

This document is available in (line-printer ready) plaintext ASCII and in PDF. It is also available in HTML from **http://www.sw.it.aoyama.ac.jp/2012/pub/draft-ietf-iri-bidi-guidelines-03.html**, and in UTF-8 plaintext from **http://www.sw.it.aoyama.ac.jp/2012/pub/draft-ietf-iri-bidi-guidelines-03.utf8.txt**. While all these versions are identical in their technical content, the HTML, PDF, and UTF-8 plaintext versions show non-Unicode characters directly. This often makes it easier to understand examples, and readers are therefore strongly advised to consult one of these versions in preference to or as a supplement to the ASCII version.

This version of this document contains bidirectional examples. In order to correctly understand the examples, it is important to view this document with a viewer that correctly implements the Unicode Bidirectional Algorithm **[UNI9]**. Many text viewers and text editors, and all major browsers, currently implement the Unicode Bidirectional Algorithm. Also, all users who are reading RTL text on a regular basis have viewers that implement this algorithm, because otherwise, they would be unable to read even the simplest texts. In order to check whether a viewer implements the Unicode Bidirectional Algorithm, please observe the following three lines:
FEDCBA ,EDCBA ,DCBA, CBA, BA, A
ب, بت, بتث, بتثج, بتثجح, بتثجحخ
א, אב, אבג, אבגד, אבגדה, אבגדהו
The first line contains upper-case Latin letters, the second line contains Arabic letters, and the third line contains Hebrew letters. Your viewer will be okay if in all three lines, the shortest word (one character) is on the right, and the longest word (six characters) on the left, the words are getting longer and longer from right to left, and the commas are between the words, but on the right of the spaces. Otherwise, please use another viewer. In the second line, the characters in each word should all be connected, and change shape slighly on context. In the first and third line, no characters should be connected.

1.3. Notation

TOC

In this document, "Bidi Notation", abbreviated "BN" is used for the given Bidi IRI examples as follows: Lower case letters a-z stand for characters that are written with a left to right ordering (such as Latin characters), whereas upper case letters A-Z represent characters that are written right to left (such as Arabic or Hebrew characters). Numbers and symbols are the same.

In this document, the key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in **[RFC2119]**.

---

2. Logical Storage and Visual Presentation

TOC

When stored or transmitted in digital representation, Bidi IRIs MUST be in full logical

order and MUST conform to the IRI syntax rules (which includes the rules relevant to their scheme). This ensures that Bidi IRIs can be processed in the same way as other IRIs.

Bidi IRIs MUST be visually ordered by the Unicode Bidirectional Algorithm **[UNIV6]**, **[UNI9]**. Bidi IRIs MUST be rendered in the same way as they would be if they were in a left-to-right embedding.

In conformance with the Unicode Bidirectional Algorithm, embedding MAY be done in one of two ways:

1. precede the IRI with U+202A, LEFT-TO-RIGHT EMBEDDING (LRE), and follow with U+202C, POP DIRECTIONAL FORMATTING (PDF); or
2. use a higher-level protocol (e.g., the dir='ltr' attribute in HTML).

Preceding and following the Bidi IRI with U+200E, LEFT-TO-RIGHT MARK (LRM) is NOT RECOMMENDED as, there are cases where this may not be sufficient to match full left to right embedding.

There is no requirement to use embedding if the display is still the same without the embedding. For example, a Bidi IRI in a text with left-to-right base directionality (such as used for English or Cyrillic) that is preceded and followed by whitespace and strong left-to-right characters does not need an embedding. Also, a bidirectional relative IRI reference that only contains strong right-to-left characters and weak characters (such as symbols) and that starts and ends with a strong right-to-left character and appears in a text with right-to-left base directionality (such as used for Arabic or Hebrew) and is preceded and followed by whitespace and strong characters does not need an embedding.

However, implementers are RECOMMENDED to use embedding in all cases where they are not completely sure that the display behavior is unaffected without the embedding.

The Unicode Bidirectional Algorithm (**[UNI9]**, section 4.3) permits higher-level protocols to influence bidirectional rendering. Such changes by higher-level protocols MUST NOT be used if they change the rendering of IRIs.

The bidirectional formatting characters that may be used before or after the IRI to ensure correct display are not themselves part of the IRI. IRIs MUST NOT contain bidirectional formatting characters (LRM, RLM, LRE, RLE, LRO, RLO, and PDF). They affect the visual rendering of the IRI but do not appear themselves. It would therefore not be possible to input an IRI with such characters correctly.

---

3. Bidi IRI Structure

TOC

The Unicode Bidirectional Algorithm is designed for general purpose text. To make sure that it does not affect the rendering of Bidi IRIs outside of the requirements of this document, some restrictions on Bidi IRIs are necessary. These restrictions are given in terms of delimiters (structural characters, mostly punctuation such as "@", ".", ":", and "/") and components (usually consisting mostly of letters and digits).

The following syntax rules from the ABNF of **[RFC3987bis]** correspond to components for the purpose of Bidi behavior: iuserinfo, ireg-name, isegment, isegment-nz, isegment-nz-nc, ireg-name, iquery, and ifragment.

Specifications that define the syntax of any of the above components MAY divide them further and define smaller parts to be components according to this document. As an example, the restrictions of **[RFC3490]** on bidirectional domain names

correspond to treating each label of a domain name as a component for schemes with ireg-name as a domain name. Even where the components are not defined formally, it may be helpful to think about some syntax in terms of components and to apply the relevant restrictions. For example, for the usual name/value syntax in query parts, it is convenient to treat each name and each value as a component. As another example, the extensions in a resource name can be treated as separate components.

For each component, the following restrictions apply:

1. A component SHOULD NOT use both right-to-left and left-to-right characters.
2. A component using right-to-left characters SHOULD start with a right-to-left character, and end with a right-to-left character potentially followed by one or more nonspacing mark (bidi class NSM).

The above restrictions are given as "SHOULD"s, rather than as "MUST"s. For IRIs that are never presented visually, they are not relevant. However, for IRIs in general, they are very important to ensure consistent conversion between visual presentation and logical representation, in both directions.

Note:
In some components, the above restrictions may actually be strictly enforced. For example, **[RFC3490]** requires that these restrictions apply to the labels of a host name for those schemes where ireg-name is a host name. In some other components (for example, path components) following these restrictions may not be too difficult. For other components, such as parts of the query part, it may be very difficult to enforce the restrictions because the values of query parameters may be arbitrary character sequences.

If the above restrictions cannot be satisfied otherwise, the affected component can always be mapped to URI notation using the general percent-encoding of IRI components, as described in **[RFC3987bis]**. Please note that the whole component has to be mapped (see also Example 9 below).

---

4. Input of Bidi IRIs

TOC

Bidi input methods MUST generate Bidi IRIs in logical order while rendering them according to **Section 2**. During input, rendering SHOULD be updated after every new character is input to avoid end-user confusion.

---

5. Examples

TOC

This section gives examples of Bidi IRIs in Bidi Notation. It shows legal IRIs with the relationship between their logical and visual representation and explains how certain phenomena in this relationship may look strange to somebody not familiar with bidirectional behavior, but familiar to users of Arabic and Hebrew. It also shows what happens if the restrictions given in **Section 3** are not followed.

To read the bidi text in the examples, read the visual representation from left to right until you encounter a block of rtl text. Read the rtl block (including slashes and other special characters) from right to left, then continue at the next unread ltr character.

Please note that "BN" stands for "Bidi Notation", see **Notation**. AR stands for Arabic,

HE for Hebrew.

Example 1: A single component with rtl characters is inverted:
Logical representation (BN): "http://ab.CDEFGH.ij/kl/mn/op.html"
Visual representation (BN): "http://ab.HGFEDC.ij/kl/mn/op.html"
Visual representation (AR): "http://ab.تثجحخد.ij/kl/mn/op.html"
Visual representation (HE): "http://ab.גדהוזח.ij/kl/mn/op.html"
Components can be read one by one, and each component can be read in its natural direction.

Example 2: More than one consecutive component with rtl characters is inverted as a whole:
Logical representation (BN): "http://ab.CDE.FGH/ij/kl/mn/op.html"
Visual representation (BN): "http://ab.HGF.EDC/ij/kl/mn/op.html"
Visual representation (AR): "http://ab.حخد.تثج/ij/kl/mn/op.html"
Visual representation (HE): "http://ab.זח.גדה/ij/kl/mn/op.html"
A sequence of rtl components is read rtl, in the same way as a sequence of rtl words is read rtl in a bidi text.

Example 3: All components of an IRI (except for the scheme) are rtl. All rtl components are inverted overall:
Logical representation (BN): "http://AB.CD.EF/GH/IJ/KL?MN=OP;QR=ST#UV"
Visual representation (BN): "http://VU#TS=RQ;PO=NM?LK/JI/HG/FE.DC.BA"
Visual representation (AR): "http://قفك#غف=ظع;ضغط=شصر?زس/نر/خد/جح.تث.اب"
Visual representation (HE): "http://שת#קר=פץ;סע=מן?כל/טי/זח.הו.גד.אב"
The whole IRI (except the scheme) is read rtl. Delimiters between rtl components stay between the respective components; delimiters between ltr and rtl components don't move.

Example 4: Each of several sequences of rtl components is inverted on its own:
Logical representation (BN): "http://AB.CD.ef/gh/IJ/KL.html"
Visual representation (BN): "http://DC.BA.ef/gh/LK/JI.html"
Visual representation (AR): "http://تث.اب.ef/gh/زس/نر.html"
Visual representation (HE): "http://אב.גד.ef/gh/כל/טי.html"
Each sequence of rtl components is read rtl, in the same way as each sequence of rtl words in an ltr text is read rtl.

Example 5: Example 2, applied to components of different kinds:
Logical representation (BN): "http://ab.cd.EF/GH/ij/kl.html"
Visual representation (BN): "http://ab.cd.HG/FE/ij/kl.html"
Visual representation (AR): "http://ab.cd.خد/جح/ij/kl.html"
Visual representation (HE): "http://ab.cd.זח/הו/ij/kl.html"
The inversion of the domain name label and the path component may be unexpected, but it is consistent with other bidi behavior. For reassurance that the domain component really is "ab.cd.EF", it may be helpful to read aloud the visual representation following the Unicode Bidirectional Algorithm. After "http://ab.cd." one reads the RTL block "E-F-slash-G-H", which corresponds to the logical representation.

Example 6: Same as Example 5, with more rtl components:
Logical representation (BN): "http://ab.CD.EF/GH/IJ/kl.html"
Visual representation (BN): "http://ab.JI/HG/FE.DC/kl.html"
Visual representation (AR): "http://ab.تث.جح/خد/نر/kl.html"
Visual representation (HE): "http://ab.גד.הו/זח/טי/kl.html"
The inversion of the domain name labels and the path components may be easier to identify because the delimiters also move.

Example 7: A single rtl component includes digits:
Logical representation (BN): "http://ab.CDE123FGH.ij/kl/mn/op.html"
Visual representation (BN): "http://ab.HGF123EDC.ij/kl/mn/op.html"
Visual representation (AR): "http://ab.حخد123تثج.ij/kl/mn/op.html"
Visual representation (HE): "http://ab.זח123גדה.ij/kl/mn/op.html"

Numbers are written ltr in all cases but are treated as an additional embedding inside a run of rtl characters. This is completely consistent with usual bidirectional text.

Example 8 (not allowed): Numbers are at the start or end of an rtl component:
Logical representation (BN): "http://ab.cd.ef/GH1/2IJ/KL.html"
Visual representation (BN): "http://ab.cd.ef/LK/JI1/2HG.html"
Visual representation (AR): "http://ab.cd.ef/زس/نر1/2خد.html"
Visual representation (HE): "http://ab.cd.ef/כל/טי1/2נר.html"
The sequence "1/2" is interpreted by the Bidirectional Algorithm as a fraction, fragmenting the components and leading to confusion. There are other characters that are interpreted in a special way close to numbers; in particular, "+", "-", "#", "$", "%", ",", ".", and ":".

Example 9 (not allowed): The numbers in the previous example are percent-encoded:
Logical representation (BN): "http://ab.cd.ef/GH%31/%32IJ/KL.html"
Visual representation (BN): "http://ab.cd.ef/LK/JI%32/%31HG.html"
Visual representation (AR): "http://ab.cd.ef/زس/نر32%/31%خد.html"
Visual representation (HE): "http://ab.cd.ef/כל/טי%32/%31נר.html"

Example 10 (allowed but not recommended):
Logical representation (BN): "http://ab.CDEFGH.123/kl/mn/op.html"
Visual representation (BN): "http://ab.123.HGFEDC/kl/mn/op.html"
Visual representation (AR): "http://ab.123.تثجحخد/kl/mn/op.html"
Visual representation (HE): "http://ab.123.nגדהוזח/kl/mn/op.html"
Components consisting of only numbers are allowed (it would be rather difficult to prohibit them), but these may interact with adjacent RTL components in ways that are not easy to predict.

Example 11 (allowed but not recommended):
Logical representation (BN): "http://ab.CDEFGH.123ij/kl/mn/op.html"
Visual representation (BN): "http://ab.123.HGFEDCij/kl/mn/op.html"
Visual representation (AR): "http://ab.123.تثجحخدij/kl/mn/op.html"
Visual representation (HE): "http://ab.123.nגדהוזחij/kl/mn/op.html"
Components consisting of numbers and left-to-right characters are allowed, but these may interact with adjacent RTL components in ways that are not easy to predict.

---

6. IANA Considerations

TOC

This document makes no changes to IANA registries.

---

7. Security Considerations

TOC

Confusion can occur with bidirectional IRIs, if the restrictions in **Section 3** are not followed. The same visual representation may be interpreted as different logical representations, and vice versa. It is also very important that a correct Unicode bidirectional implementation be used.

---

8. Acknowledgements

TOC

This document was derived from **[RFC3987]** and **[RFC3987bis]** and the acknowledgments of those documents apply. Shunsuke Oshima (大嶋 俊介) provided the data for **Appendix A**.

9. Main Changes Since RFC 3987

TOC

This section describes the main changes since **[RFC3987]**.

- Separated out the section on bidi in **[RFC3987]** to this document.
- Added examples in Arabic and Hebrew, which can be seen in html/pdf/utf8.txt versions.
- Allowed NSMs at the end of components, for Dhivehi, Yiddish,...
- TODO: check for major changes between RFC3987 and draft -02.

Note to RFC Editor: Please remove this paragraph before publication. Detailled change logs are available in the IETF tools subversion repository at http://trac.tools.ietf.org/wg/iri/trac/log/draft-ietf-iri-3987bis/draft-ietf-iri-bidi-guidelines.xml.

10. References

TOC

10.1. Normative References

TOC

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels," BCP 14, RFC 2119, March 1997 (TXT, HTML, XML).

[RFC3490] Fältström, P., Hoffman, P., and A. Costello, "Internationalizing Domain Names in Applications (IDNA)," RFC 3490, March 2003 (TXT).

[RFC3987bis] Dürst, M., Masinter, L., and M. Suignard, "Internationalized Resource Identifiers (IRIs)," October 2012.

[UNI9] Davis, M., "The Unicode Bidirectional Algorithm," Unicode Standard Annex #9, September 2012.

[UNIV6] The Unicode Consortium, "The Unicode Standard, Version 6.2.0 (Mountain View, CA, The Unicode Consortium, 2012, ISBN 978-1-936213-07-8)," October 2012.

10.2. Informative References

TOC

[RFC3987] Dürst, M. and M. Suignard, "Internationalized Resource Identifiers (IRIs)," RFC 3987, January 2005 (TXT).

Appendix A. List of ASCII Symbols and their Bidirectional Character Types

TOC

To help understand the influence of various symbols on IRI display, this appendix lists all of them, giving the character itself, the Unicode codepoint, the character name, the bidirectional character type (BCT) and the rule and relevance in the IRI syntax.

The most important ones in practice are ":", delimining schem and port (CS, Common Number Separator), "/" to indicate generic (hierarchical) schemes and as a path separator (CS, Common Number Separator), "?" to introduce a query part (ON, Other Neutral), "#" to introduce a fragment identifier (ET, European Number Terminator), "." to separate labels in a domain name (CS, Common Number Separator), "&" to separate form parameters (ON, Other Neutral), and "@" to separate user information (ON, Other Neutral).

```
Char Codepoint  Character Name       BCT  IRI syntax
-------------------------------------------------------------
```

"#"	U+0023	NUMBER SIGN	ET	gen-delims, fragments
"/"	U+002F	SOLIDUS	CS	gen-delims, paths
":"	U+003A	COLON	CS	gen-delims, scheme, port
"?"	U+003F	QUESTION MARK	ON	gen-delims, query part
"@"	U+0040	COMMERCIAL AT	ON	gen-delims, user
"["	U+005B	LEFT SQUARE BRACKET	ON	gen-delims
"]"	U+005D	RIGHT SQUARE BRACKET	ON	gen-delims
"%"	U+0025	PERCENT SIGN	ET	pcd-encoded
"!"	U+0021	EXCLAMATION MARK	ON	sub-delims
","	U+002C	COMMA	CS	sub-delims
"+"	U+002B	PLUS SIGN	ES	sub-delims
"$"	U+0024	DOLLAR SIGN	ET	sub-delims
"("	U+0028	LEFT PARENTHESIS	ON	sub-delims
"'"	U+0027	APOSTROPHE	ON	sub-delims
")"	U+0029	RIGHT PARENTHESIS	ON	sub-delims
"*"	U+002A	ASTERISK	ON	sub-delims
";"	U+003B	SEMICOLON	ON	sub-delims
"="	U+003D	EQUALS SIGN	ON	sub-delims, forms
"&"	U+0026	AMPERSAND	ON	sub-delims, forms
"."	U+002E	FULL STOP	CS	unreserved, domain names
"-"	U+002D	HYPHEN-MINUS	ES	unreserved
"_"	U+005F	LOW LINE	ON	unreserved
"~"	U+007E	TILDE	ON	unreserved
" "	U+0020	SPACE	WS	excluded, delim
'"'	U+0022	QUOTATION MARK	ON	excluded, delim
"\"	U+005C	REVERSE SOLIDUS	ON	excluded, unwise
"^"	U+005E	CIRCUMFLEX ACCENT	ON	excluded, unwise
"<"	U+003C	LESS-THAN SIGN	ON	excluded, delim
">"	U+003E	GREATER-THAN SIGN	ON	excluded, delim
"`"	U+0060	GRAVE ACCENT	ON	excluded, unwise
"|"	U+007C	VERTICAL LINE	ON	excluded, unwise
"{"	U+007B	LEFT CURLY BRACKET	ON	excluded, delim
"}"	U+007D	RIGHT CURLY BRACKET	ON	excluded, delim

Authors' Addresses

TOC

Martin J. Dürst
Aoyama Gakuin University (青山学院大学)
5-10-1 Fuchinobe
Chuo-ku
Sagamihara, Kanagawa 252-5258
Japan
Phone: +81 42 759 6329
Fax: +81 42 759 6495
Email: duerst@it.aoyama.ac.jp
URI: http://www.sw.it.aoyama.ac.jp/Dürst/

Larry Masinter
Adobe
345 Park Ave
San Jose, CA 95110
U.S.A.
Phone: +1-408-536-3024
Email: masinter@adobe.com
URI: http://larry.masinter.net

Adil Allawi (عادل علاوي)
Diwan Software Limited
37-39 Peckham Road
London SE5 8UH
United Kingdom
Phone: +44 7718 785850
Fax: +44 20 72525444
Email: adil@diwan.com
URI: http://ironymark.diwan.com/