The Unicode Consortium Discussion Forum

Forum rules

Please click here to view the forum rules

Discussion of PRI 185

Author	Message
mark	**Post subject:** Discussion of PRI 185 **Posted:** Sat Sep 24, 2011 9:27 am

offline

Forum Admin

Joined: Fri Dec 04, 2009 6:13 pm
Posts: 28

The Unicode Bidirectional Algorithm (UBA), specified in Unicode Standard Annex #9, was designed for handling ordinary text, and predated the rise of the web. Unfortunately, IRI/URLs are not ordinary text; they are syntactically complex in ways that don't work well with the UBA. That causes IRIs that contain right-to-left text (such as Arabic or Hebrew) to appear jumbled, to the point where the IRIs are either uninterpretable, misleading, or ambiguous. In particular the ambiguous displays could cause security problems.

The background document for this PRI provides a detailed description of the problem, and proposes a solution. The Unicode Technical Committee would like feedback on the feasibility of the proposal, and in particular, on the open issues listed in the background document.

For more information, see http://www.unicode.org/review/pri185/

Note: This is a moderated forum, so as to allow for more discussion of the issues. Please file comments on PRI 185 here, rather than on the Unicode mailing lists.

To make the discussion more effective, please:

* Stay on topic; avoid subjects not directly relevant to PRI 185
* Keep messages short and understandable
* Try to focus on only one issue per message.

Top profile pm email edit quote

matial

offline

Joined: Sun Aug 22, 2010 2:14 am
Posts: 4

Post subject: Re: Discussion of PRI 185 **Posted:** Sun Oct 23, 2011 8:00 am

My personal preference goes to Option 1 "Constant Order".

However, in discussions at SII (the Standards Institution of Israel), including feedback solicited from the public, there were voices to sustain that such a standard would not be implemented, and that URLs should be displayed in a consistent RTL order if they contain **only** Hebrew.
I think that such a statement must be interpreted as follows:
1. "Only Hebrew" in fact means at least one Hebrew letter, optional digits and neutrals, and no LTR characters.
2. Only the domain part is to be considered in the determination of the display order. The order of fields in path, query and fragment will be set identical to the order determined for the domain.
3. It is not clear if the scheme must be considered or not. If yes, this proposal will give the same results as constant LTR order until some schemes are defined in Hebrew letters.

Based on the principles above, the URL display direction would be determined as LTR or RTL and all fields (according to the definition of fields in the background document in section "Proposed extension of UBA for bidi_IRIs") must be laid out one after the other according to the URL display direction, independently of the direction of the context (paragraph).

Please add this option to those considered in the background document.

Note that the consensus at SII for file names is still to use Option 1 "Constant Order", which is in fact constant **LTR** order.

I don't remember an explicit discussion of mail addresses, I assume that the rules for mail addresses should follow the rules for URLs.

_____
Matitiahu Allouche
bidi user, Israel

Top profile pm email edit quote

mark

offline

Forum Admin

Joined: Fri Dec 04, 2009 6:13 pm
Posts: 28

Post subject: Re: Discussion of PRI 185 **Posted:** Sun Oct 23, 2011 10:20 am

That sounds like an interesting option, and could apply to other RTL scripts as well. It might be a workable compromise.

One question is whether it should be only the domain name, or should also include the path. While the query and fragment part pale in comparison, the path is different.

The distinction between which components are in the path, and which in the domain name is pretty arbitrary. Consider, for example:

http://www.economist.com/world/europe

vs the perfectly reasonable alternative choices that The Economist could have chosen:

http://world.economist.com/europe, or
http://europe.world.economist.com

Top | profile | pm | email | edit | quote

Jony

offline

Joined: Tue Jan 11, 2011 8:23 pm
Posts: 1

Post subject: Re: Discussion of PRI 185 **Posted:** Sun Oct 23, 2011 10:49 am

I don't see any reason to change my previously stated position. Don't tinker with the UBA. Stability is critical. The justification for the proposal is insufficient: in a few years RTL URLs will be purely RTL, some countries are moving faster, some don't yet see the need; the problem of mixed directionality URLs will be gone and we will be left with the proposed added complexity.

Top | profile | pm | email | edit | quote

mark

offline

Forum Admin

Joined: Fri Dec 04, 2009 6:13 pm
Posts: 28

Post subject: Re: Discussion of PRI 185 **Posted:** Sun Oct 23, 2011 11:30 am

What the UTC is looking at is an extension of the UBA, not a change. That is, UBA conformance would not require implementation of the extension. People could implement the plain UBA, or *in addition* claim conformance to the extension. The higher level protocol clauses of the UBA permit such extensions.

The reason for the UTC to standardize on any such extension is that we are seeing a great deal of demand for handling BIDI IRIs. The concern is that if nothing is done, we'll end up with arbitrary differences between different vendors' implementations of BIDI IRI extensions, which would be a huge mess for everyone.

(Personally, I would like to see technology changes that would allow any server to easily support IRIs without any LTR characters, so that the security and usability problems would be minimized. However, many people believe that that is infeasible in practice: that it would be years or never before implementations would be able to support RTL schemes and extensions, such as index.pdf.)

Top | profile | pm | email | edit | quote

amitar

offline

Joined: Wed Dec 08, 2010 4:21 am
Posts: 2

Post subject: Re: Discussion of PRI 185 **Posted:** Sun Oct 23, 2011 3:17 pm

> **Jony wrote:**
>
> in a few years RTL URLs will be purely RTL ... the problem of mixed directionality URLs will be gone and we will be left with the proposed added complexity.

One should keep in mind that the problems that this extension tries to address are not limited to mixed directionality IRI (though they do constitute the most common instances). For example, numerical path elements within a "fully RTL" IRI might are displayed in a very confusing order if you use the unmodified UBA.
Memory order "AB/C1/234/DEF" is displayed "FED/1/234C/BA", while the correct display (the one matching the semantics of the path) should be "FED/234/1C/BA" (which is supported by options 2-4 and by matial's comment).

The relevance of the mixed directionality cases depends on estimates for the time between adoption of the proposal to (almost) global adoption of RTL IRI's. It seems likely that this period would be long enough for relevance (but of course I might be wrong).
I tend to support option 4 (deduce directionality from TLD).

However, I have doubts regarding the proposed IRI recognition scheme: First, it does not handle filenames (I doubt they could be safely auto-recognized in plaintext at all).
Second, while the algorithm seems fine for today's commonly used cases, it depends on specific data lists (such as TLD) which might change over time.

Top profile pm email edit quote

shai

offline

Joined: Sun Oct 23, 2011 4:02 pm
Posts: 1

Post subject: Re: Discussion of PRI 185 **Posted:** Sun Oct 23, 2011 10:31 pm

I would like to comment on a few issues:

1) I completely agree with amitar on the issue of IRI detection in plain text. I further submit that IRI detection is context-dependent, and therefore can usually be done better by specific applications, where assumptions on usage can be made, than in the general case.

2) As a partial solution to the problem in plain text, I propose the well-known Unicode mechanism of control characters; just like we have LRE..PDF, add a control character, say IRI, so that any text in IRI..PDF will have its BiDi properties interpreted according to IRI rules (and if the text cannot be parsed as an IRI, assume it is a file name).

This suggestion has one outstanding disadvantage, in that it makes the proposed modification an amendment rather than an extension; that is, I can't see how in the presence of such a control character, conformance can be kept optional.

3) I am not decided about the options for field ordering. My favorites (in no particular order) are

- matial's suggestion above: "domain order", where the domain dictates RTL order if and only it is has no strong LTR characters and at least one strong RTL character; and
- A suggestion made by Shachar Shemesh in SII discussions: If the IRI is all RTL, digits or neutrals, make it all RTL; otherwise, use constant LTR ordering, but consider the domain as one composite field with its sub-fields ordered according to the direction of the TLD, and the TLD

itself last among them.
At any rate, I think that if the scheme is stated explicitly, it should set the direction for the entire IRI (whether the domain is considered as one composite field or a set of fields).

Thus, with matial's suggestion, where the memory order is

Code:

```
AB.CD.COM/EF/GH
```

We have according to either proposal a display order of

Code:

```
HG\FE\MOC.DC.BA
```

but with matial's we get

Code:

```
http://BA.DC.MOC/FE/GH
```

and with Shachar's

Code:

```
http://MOC.DC.BA/FE/GH
```

Top profile pm email edit quote

doron

offline

Joined: Sun Oct 23, 2011 8:53 am
Posts: 1

Post subject: Re: Discussion of PRI 185 **Posted:** Mon Oct 24, 2011 3:42 pm

I'd like to first echo matial's proposal (and further discussion by shai), I think it would have made the most sense in terms of usability and readability.

I'm saying would have made, because I think the main problem here is the plain text detection part. The proposal presumes a preloaded, maintained list of TLDs as the basis for such detection.
Apparently, this will quite soon become quite unmanageable. Recently, ICANN has announced a plan to add a large number of gTLDs to the root DNS zone. The exact rate of new TLDs / year is yet to be seen, but the statement was, IIRC, *up to* 1,000 per year.

Without commenting on the sensibility of this step, it would, seemingly, make any plan to use a fixed (or semi fixed) set of TLDs as the anchor for plain text detection of IRIs quite unworkable.

Top profile pm email edit quote

efratian

offline

Joined: Sat Sep 24, 2011 10:47 pm
Posts: 1

Post subject: Re: Discussion of PRI 185 **Posted:** Wed Oct 26, 2011 6:27 am

1. I think it would be best not to attempt to standardize an algorithm for detecting IRIs (and email addresses, file paths) in plain text.

However, I do think it is important to provide a standard algorithm for determining the visual ordering of a string *that is already known to be* an URI (that would give better results than the UBA without a higher level protocol).

Applications that know that they are displaying an URI (or email address or file path) can call this algorithm before displaying the string. The obvious "customers" would be browsers (e.g. in the address bar), email UIs (when displaying an email address), file system viewers (when display a path). Furthermore, the many applications that already look for URIs and email addresses in plain text and "linkify" them can easily start applying this algorithm for their display - without having to modify their logic for detecting URIs and email addresses in plain text, which may be a much more contentious issue.

The advantages of this approach are:
- No change or even extension to the UBA. Just an additional algorithm for a higher-level protocol that applications can apply on top of the UBA when displaying a string that they already know to be a URI, email address, etc.
- Unicode can stay away from the issue of detecting IRIs in plain text, for which there is no complete solution.
- Applications that already look for URIs in plain text do not have to change their logic for that to what they may feel is a too-complicated, too-expensive, or less-good algorithm that they already have.
- An application, e.g. the browser address bar, can force the display of *any* string it knows to be an URI as an URI. There is no chance of a spoofer concocting a URI that gets displayed in a misleading way even in the browser address bar by intentionally falling outside the syntax for detecting URIs in plain text, whatever it may be.

The disadvantage is that bidi IRIs in plain text will continue to be displayed poorly when the application displaying the plain text does not look for URIs. I do not think that this is a big problem, for two reasons:
- Once the user (laboriously) pastes the IRI into the browser address bar, the IRI will be displayed correctly. The user has a chance to detect foul play.
- Even if the UBA itself were modified today to look for IRIs in plain text, it would take years for the change to make its way into every last application that displays plain text.

2. I think that getting IRI (and email address, file path) components displayed in a consistent direction is important. The details of the algorithm for choosing the direction are less important than that some algorithm be decided upon and implemented.

3. I do not think that there is a perfect way to choose the overall direction.

The domain direction criterion (i.e. RTL if the domain contains at least one R or AL character and no L characters) allows IRIs to remain ambiguous. For example, both "http://SITE.RTLD?com.anothersite//:http" (in logical order, where uppercase Latin letters are meant to represent RTL letters) and "http://anothersite.com?RTLD.SITE //:http" would be displayed as "http://anothersite.com?DLTR.ETIS//:http". This is an obvious opportunity for spoofing.

The always-LTR criterion means that all-RTL IRIs are very difficult to read: the eyes have to go backwards and forwards again and again.

The context direction criterion means that the vast majority of existing, all-LTR URIs would be very difficult to read in an RTL context: the eyes would have to go backwards and forwards again and again. Furthermore, the sudden switch from the familiar http://www.myfavoritesite.com to the bizarre-looking com.myfavoritesite.www://http would be very difficult on users. Furthermore, the use of a or LRE/RLE+PDF around the IRI would be an easy way to spoof.

Etc.

Given this choice, my preference is for the always-LTR criterion. If that is impossible, the domain direction criterion would be my next choice.

_____
Aharon Lanin

Top profile pm email edit quote

verdyp

offline

Joined: Fri Sep 23, 2011 9:56 am
Posts: 7
Location: Niort, France

Post subject: Re: Discussion of PRI 185 **Posted:** Sat Oct 29, 2011 9:40 pm

I also strongly favor the fixed display order of fields in URI and IRI. Only individual labels (in the domain name part), or path entries (separated by /) or parameters (separated by ? = &) should apply the UBA individually.

However the situation is more complex when URI and IRI are appearing as part of a text content where the UBA algorithm is applied globally). That's why I strongly suggest a separate parsing and display when extracting an URI/IRI from a larger text, without automatically creating an activable link (this step should require a confirmation for the extracted URI/IRI, to use the more restrictive UBA parsing, field by field: field separators will always be ouside these fields, and fields will be displayed in the confirmed extracted URI in the fixed order).

The URI/IRI parser should then be able to parse the domain name, checking where there's a difference of DNS delegation between a parent domain (including a TLD) and a subdomain.

Unfortunately, the way most browsers are working is that they don't check from which authoritative DNS server the resolution will come from (they just perform a basic query with type "A" or "A+AAAA" to convert the domain into an IPv4 or IPv6 address, but don't request the "SOA" entry and dns entries; most of these queries are then made directly to the upstream caching DNS server of the local ISP, which deliveres non authorititive cached entries, but which also does not perform the anti-spoofing checks and which are also not prepared to scale for delivering replies to massive SOA checks: if we request to everyone to perform SOA checks, the whole DNS system would collapse from the root and TLD domain levels, as the DNS system only scales because of the presence of caching DNS servers everywhere; the only solution would be that non-authoritative replies coming from caches would contain a digital signature from the parent domain, for each answer to any query type).

So instead of just requesting "A" or "A+AAAA" you would request "A+DSIG" or "A+AAAA+DSIG" but this would increase the data volume of the replies with these signatures delivered by caches... This would require that the parent domain delivering those digital signatures also deliver their own siganture certificates in their SOA records (for example by querying the parent domain for its "CERT" entry, that can itself be signed by its parent domain). Once problem is that delivering a digital signature will still work with "A+DSIG" or "A+AAAA+DISG" requests via UDP, but certificates are generally much larger than a single datagram, and can only be delivered from DNS servers via TCP (and then there's a problem for the number of TCP port numbers available to all requesters of a DNS server: querying a DNS server with TCP is generally much more restricted in terms of number/frequency of queries allowed per client, than with UDP). For now, the worldwide DNS system is not prepared for massive delivery of certificates by TCP, even via caching DNS servers (it would probably work easily with the root and TLD domains, that can easily be cached, but not for the many generic second level domains hosted by registrars, if they each need their own signing certificate for their subdomains including "www").

For now, there's such a specification of signature schemes (DNSSEC), but its deployment is still late (only realized now in the root zone, and in a few gTLD like ".com", but not in most other TLDs). And the use of DNSSEC for queries is still limited to secure protocols, but most DNS queries are unsecured (most caching DNS servers cannot currently scale for its massive general use; the query time is still very long and the number/frequency of queries remains very limited).

Top profile pm email edit quote

matial

offline

Joined: Sun Aug 22, 2010 2:14 am
Posts: 4

Post subject: Re: Discussion of PRI 185 **Posted:** Sun Oct 30, 2011 7:54 am

I hope this post is still in time to be considered before the coming UTC meeting.

I think that we should distinguish 2 issues in PRI 185:
a) How should an IRI, an address mail or a file name containing RTL characters be displayed?
b) How can an IRI, an address mail or a file name be identified within plain text?

About item b, I think it should be left outside the UBA. It is up to applications to locate occurrences of IRIs etc...
After locating such an occurrence, the application should invoke a special version or enhanced mode of the UBA to display the occurrence according to the desired format.
UTC can propose an algorithm to identify IRIs etc..., but this algorithm should not be normative, and it should not be part of the UBA.

About item a, with all respect due to UTC, I think that this is not a technical decision to be made by Unicode experts but a usability issue to be solved by representatives of the respective user groups. I know by experience that it is difficult to find a relevant sample of the user community. Maybe this should be answered by the respective national bureaus. Anyway, I don't think that UTC can statute on that.
After a conclusion is achieved about the proper representation of IRIs etc..., an extension to the UBA should be formulated to handle them, and this extension should be normative.

_____
Matitiahu Allouche
bidi user, Israel

Top profile pm email edit quote

Display posts from previous: All posts Sort by Post time Ascending Go

newtopic postreply **Page 1 of 1** [11 posts]

Board index » Public Review Discussions » PRI 185 - Extension of UBA for improved display of URL/IRIs
All times are UTC - 8 hours

Who is online

Users browsing this forum: **Sarasvati** and 0 guests

Quick-mod tools: Lock topic Go

You **can** post new topics in this forum
You **can** reply to topics in this forum
You **can** edit your posts in this forum
You **can** delete your posts in this forum
You **can** post attachments in this forum

Search for: Go

Jump to: PRI 185 - Extension of UBA for improved display of URL/IRIs Go

[Administration Control Panel]