



The Unicode Consortium Discussion Forum

[Forum Home](#)
[Unicode Home Page](#)
[Code Charts](#)
[Technical Reports](#)
[FAQ Pages](#)
[Forum FAQ](#) - [Smartfeed](#) - [0 new messages](#) - [Search](#) - [Members](#) - [User Control Panel](#) - [Logout](#) [[Sarasvati](#)]

Last visit was: Mon Oct 31, 2011 8:36 pm

It is currently Tue Nov 01, 2011 9:46 am

[View unanswered posts](#) | [View active topics](#)
[View new posts](#) | [View your posts](#)
[Board index](#) » [Public Review Discussions](#) » [PRI 185 - Extension of UBA for improved display of URL/IRIs](#)

All times are UTC - 8 hours

Forum rules

Please click here to view the forum rules

weak recognition of domain names as IRIs

[\[Moderator Control Panel \]](#)


Page 1 of 1 [2 posts]

[Subscribe topic](#) | [Bookmark topic](#) | [Print view](#) | [E-mail friend](#) [Previous topic](#) | [First unread post](#) | [Next topic](#)

Author

Message

verdyp

Post subject: weak recognition of domain names as IRIs

 Posted: Sat Sep 24, 2011 9:44 am

Joined: Fri Sep 23, 2011 9:56 am

Posts: 7

Location: Niort, France

Another issue will soon appear with the proposal: to recognize an IRI, an application will likely have to recognize either IRI starting with a known scheme (which is something quite stable), or recognize IRI that are much more weakly indicated as domain names. The proposal is to recognize a list of dot-separated labels (possibly internationalized with IDNA) and whose final label is a known TLD.

However, ICANN is in the process of adding soon MANY TLD, for lots of generic terms. So we would end in having to recognize as an IRI every sequence consisting of a word followed by a dot and followed by another word, even if words are restricted to some letters, or the hyphen and ZW(N)J.

In this condition, we would recognize lots of abbreviations or acronyms as IRIs, notably if they are noted with dot separators, and the final letters occur in at least a pair (because the ICANN policy is to not encode TLDs with less than 2 letters).

Could we instead of allowing ALL possible TLDs that may appear in the future, only allow a restricted list, and only if the TLD is written in lowercase ? For example "unicode.org" or "Unicode.org" would be recognized as a simple IRI, but NOT "unicode.ORG" or "Unicode.Org" (still considered as an abbreviation or acronym, and that are not required to be recognized as domain names in plain text)...

Is there a policy at ICANN related to the lettercase of LTDs ? And what about the newly proposed international TLDs, written in caseless scripts such as Arabic, Sinograms, Kanas or Hangul ?

Now let's look at scripts that do not consistently use whitespaces to separate words. What would happen to "<Lao word>.<Lao word>" if there's a new TLD in Lao? How can we avoid them being incorrectly assumed as IRIs (and possibly blocked by tools like antivirus and antispam softwares, that would start focusing legitimate texts) ?

Can't the proposal limit more the automatic recognition of weakly specified IRI (without the leading scheme), by recommending authors that want domain names being recognized as IRI to use some canonical form, and possibly only in a restricted set of TLDs (for example ccTLDs, and gTLD using Latin), forcing them to use a better textual indication that this is a valid link ?

In my opinion, weakly defined IRIs using just a domain name should not generate a link automatically in the application rendering the text. But the application may propose a contextual menu when clicking somewhere or selecting text, that allows parsing this content and propose an editable link that can then be followed with a second click or confirmation by the user, or active selection in a list of candidate links.

This would also improve the rendering time for standard display of plain texts. The usage pattern would include parsing SMS or "Tweets", as well as standard plain-text emails, or messages posted in online forums without rich-text options or with restricted options, with a bonus in terms of security for readers (against accidental clicks when they just intended to select text for copy-paste operations like citations).



Top



verdyp

Post subject: Re: weak recognition of domain names as IRIs

Posted: Sat Oct 29, 2011 9:03 pm

offline

Joined: Fri Sep 23, 2011
9:56 am

Posts: 7

Location: Niort, France

Maybe it can be clearer if I give an example:

can you guess from the syntax if "iv.xxx" is a domain name or a pair of roman numbers separated by a dot ?

If you cannot, because there's no leading "http://" URI scheme prefix, then the user agent should not create a link that can be activated with a simple link. Instead it may provide a right-click contextual menu, that would parse the text around the position in the clicked (or selected) text element, and then propose an action menu offering to follow the link to the weakly detected link. The action menu should display the parsed link completely, possibly by opening a dialog where extra characters (notably punctuations) can be edited. This editable dialog should validate the link format before continuing with the OK button, and should still continue to evaluate the risks for the domain (such as spoofed confusable characters). Ideally it should also allow viewing the IRI in its standard URI form (punyencoded domain name, %nn encoding of the path and parameters converted to UTF-8). The mere basic display of the target link in the status bar is now not enough.

And more attention should be given if the detected link goes to another domain name or if there's a switch from a secure protocol (mostly HTTPS) to a non-secure protocol (most often HTTP). This should also be the case when going to a subdomain or parent domain of the current domain, due to the growing concern caused by abused DNS servers for the subdomains, not hosted by the domain name registrar, notably if the HTTPS certificate does not restrict its usage to only one domain but to all subdomains in the current domain. This is now becoming very serious because DNS servers used by domain registrars are now

becoming very secure, but not when many organizations start using their own local DNS servers with minimal administration and security checks: the main website is generally stable, but we can find now spurious subdomains that have been inserted by hackers trying to profit of the DNS delegation for subdomains on an private DNS server (or its web administration interface).



Top



Display posts from previous: Sort by



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:

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:

Jump to:

[Administration Control Panel]

Powered by phpBB © 2000, 2002, 2005, 2007 phpBB Group
 Template made by DEVPPPL.com