

UrlSkip

Web URL Simplification in Scheme

Version 0.1
3 January 2005

Neil W. Van Dyke
neil@neilvandyke.org

<http://www.neilvandyke.org/urlskip/>

Copyright © 2005 Neil W. Van Dyke. This program is Free Software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version. This program is distributed in the hope that it will be useful, but without any warranty; without even the implied warranty of merchantability or fitness for a particular purpose. See the GNU General Public License [GPL] for details. For other license options and commercial consulting, contact the author.

1 Introduction

The UrlSkip Scheme library provides a function that translates some of the Web URLs that might be used to track a user across sites, by removing intermediate HTTP redirectors or information that might identify the user. Such a function might be used as part of a privacy-enhancing Web browser, or to canonicalize or un-obfuscate URLs for Web analysis projects.

Note that UrlSkip is not intended to remove information used by “affiliate” referral programs to identify site operators that have sent users to a site. However, in some cases this affiliate ID information might be lost in the process of removing a intermediary URL that is used by a third party to track and profile users.

UrlSkip currently requires R5RS, the `[uri.scm]` library, and a particular regular expression function. Therefore, UrlSkip currently works only with PLT MzScheme, although it will be made more portable once `uri.scm` is.

UrlSkip is released under the GPL license, unlike most of the author’s other released Scheme libraries, which are LGPL.

2 Host Handlers

The procedures in this section are used internally by the `urlskip` procedure, and correspond to particular HTTP server hostnames. They are exposed here mainly for purposes of documentation, and are likely to change in future versions of UrlSkip. Each procedure accepts a *uriobj* and yields either a new URL string of a simpler URL, or `#f` if no simpler URL was determined.

<code>urlskip-http-ad-doubleclick-net</code> <i>uriobj</i>	[Procedure]
UrlSkips <code>http://ad.doubleclick.net:</code> Substring following <code>;;~sscs=%3f</code> .	
<code>urlskip-http-click-linksynergy-com</code> <i>uriobj</i>	[Procedure]
UrlSkips <code>http://click.linksynergy.com:</code> If path <code>/fs-bin/stat</code> , then query value <code>RD_PARM1</code> or <code>rd_parm1</code> .	
<code>urlskip-http-rds-yahoo-com</code> <i>uriobj</i>	[Procedure]
UrlSkips <code>http://rds.yahoo.com:</code> Substring of the <code>http</code> URL following <code>*-</code> .	
<code>urlskip-http-service-netmeans-com</code> <i>uriobj</i>	[Procedure]
UrlSkips <code>http://service.netmeans.com:</code> If path <code>/bfast/click</code> , then query value <code>loc</code> .	
<code>urlskip-http-web-ask-com</code> <i>uriobj</i>	[Procedure]
UrlSkips <code>http://web.ask.com:</code> If path <code>/redir</code> , then query value <code>bu</code> .	
<code>urlskip-http-www-amazon-com</code> <i>uriobj</i>	[Procedure]
UrlSkips <code>http://www.amazon.com:</code> If path <code>/exec/obidos/redirect</code> , then remove all query values except for <code>tag</code> and <code>path</code> .	
<code>urlskip-http-www-anrdoezrs-net</code> <i>uriobj</i>	[Procedure]
UrlSkips <code>http://www.anrdoezrs.net:</code> Query value <code>url</code> .	
<code>urlskip-http-www-commission-junction-com</code> <i>uriobj</i>	[Procedure]
UrlSkips <code>http://www.commission-junction.com:</code> If path <code>/track/track.dll</code> , then query value <code>URL</code> .	
<code>urlskip-http-www-google-com</code> <i>uriobj</i>	[Procedure]
UrlSkips <code>http://www.google.com:</code> If path <code>/pagead/iclk</code> , then query value <code>adurl</code> . If path <code>/url</code> , then query value <code>q</code> .	
<code>urlskip-http-www-qksrv-net</code> <i>uriobj</i>	[Procedure]
UrlSkips <code>http://www.qksrv.net:</code> Query value <code>loc</code> or <code>url</code> .	

3 Interface

The only real library interface is the `urlskip` procedure.

<code>urlskip</code> <i>uri</i>	[Procedure]
Accepts a URL <i>uri</i> and yields a URL that is either <i>uri</i> or a <code>UrlSkip</code> simplified version of same. <i>uri</i> may be a string or a <i>uriobj</i> . If a simplified URL is yielded, it is always a string.	

4 Tests

The UrlSkip test suite can be enabled by editing the source code file and loading [Testeez]; the test suite is disabled by default.

History

Version 0.1 — 3 January 2005
Initial release.

References

- [GPL] Free Software Foundation, “GNU Lesser General Public License,” Version 2, June 1991, 59 Temple Place, Suite 330, Boston, MA 02111-1307 USA.
<http://www.gnu.org/copyleft/gpl.html>
- [Testeez] Neil W. Van Dyke, “Testeez: Simple Test Mechanism for Scheme,” Version 0.1.
<http://www.neilvandyke.org/testeez/>
- [uri.scm] Neil W. Van Dyke, “uri.scm: Web Uniform Resource Identifiers (URI) in Scheme,” Version 0.1.
<http://www.neilvandyke.org/uri-scm/>