Ruby - Bug #19266

URI::Generic should use URI::RFC3986_PARSER instead of URI::DEFAULT_PARSER

12/26/2022 06:52 PM - gareth (Gareth Adams)

Status:	Closed		
Priority:	Normal		
Assignee:	hsbt (Hiroshi SHIBATA)		
Target version:			
ruby -v:	ruby 3.1.3p185 (2022-11-24 revision 1a6b16756e) [arm64-darwin21]	**Backport:**	2.7: UNKNOWN, 3.0: UNKNOWN, 3.1: UNKNOWN, 3.2: UNKNOWN

Description

In June 2014, uri/common was updated to introduce a RFC3986-compliant parser (URI::RFC3986_PARSER) as an alternative to the previous RFC2396 parser, and common methods like URI() were updated to use that new parser by default. The only methods in common not updated were URI.extract and URI.regexp which are marked as obsolete. (The old parser was kept in the DEFAULT_PARSER constant despite it not being the default for those methods, presumably for backward compatibility.)

However, similar methods called on URI::Generic were never updated to use this new parser. This means that methods like URI::Generic.build fail when given input that succeeds normally, and this also affects subclasses like URI::HTTP:

```
$ pry -r uri -r uri/common -r uri/generic

[1] pry(main)> URI::Generic.build(host: "underscore_host.example")
URI::InvalidComponentError: bad component(expected host component): underscore_host.example
from /Users/gareth/.asdf/installs/ruby/3.1.3/lib/ruby/3.1.0/uri/generic.rb:591:in `check_host'

[2] pry(main)> URI::HTTP.build(host: "underscore_host.example")
URI::InvalidComponentError: bad component(expected host component): underscore_host.example
from /Users/gareth/.asdf/installs/ruby/3.1.3/lib/ruby/3.1.0/uri/generic.rb:591:in `check_host'

[3] pry(main)> URI("http://underscore_host.example")
=> #<URI::HTTP http://underscore_host.example>
```

URI::Generic.new allows a configurable parser positional argument to override the class' default parser, but other factory methods like .build don't allow this override.

Arguably this doesn't cause problems because at least in the case above, the URI can be built with the polymorphic constructor, but having the option to build URIs from explicit named parts is useful, and leaving the outdated functionality in the Generic class is ambiguous. It's possible that the whole Generic class and its subclasses aren't intended to be used directly how I'm intending here, but there's nothing I could see that suggested this is the case.

I'm not aware of the entire list of differences between RFC2396 and RFC3986. The relevant difference here is that in RFC2396 an individual segment of a host (domainlabels) could only be alphanum | alphanum *(alphanum | "-") alphanum, whereas RFC3986 allows hostnames to include any of ALPHA / DIGIT / "-" / "." / "_" / "~". It's possible that other differences might cause issues for developers, but since this has gone over 8 years without anyone else caring about this, this is definitely not especially urgent.

Related issues:

Related to Ruby - Bug #19756: URI::HTTP.build does not accept a host of `_gat...	**Open**

Associated revisions

Revision f76a4cda86afef8f9007b403febeb5524269f007 - 12/04/2024 05:34 AM - hsbt (Hiroshi SHIBATA)

Added Bug #19266, Bug #20795 and net-http changes about removing deprecated constants to NEWS

Revision f76a4cda86afef8f9007b403febeb5524269f007 - 12/04/2024 05:34 AM - hsbt (Hiroshi SHIBATA)

Added Bug #19266, Bug #20795 and net-http changes about removing deprecated constants to NEWS

Revision f76a4cda - 12/04/2024 05:34 AM - hsbt (Hiroshi SHIBATA)

Added Bug #19266, Bug #20795 and net-http changes about removing deprecated constants to NEWS

History

#1 - 01/10/2023 05:27 AM - gareth (Gareth Adams)

- File 0001-Update-URI-Generic.build-build2-to-use-RFC3986_PARSE.patch added

The attached patch adds a failing test and a change that fixes it.

The rest of the test suite passes with this patch.

#2 - 01/16/2023 05:09 PM - gareth (Gareth Adams)

After a couple of weeks with no reply I wanted to ask if I could get at least one comment on this issue?

A quick summary of the issue:

- In 2014, URI was updated to use a new RFC3986-compliant parser by default instead of the previous RFC2396 parser.
- Two methods inside URI::Generic (build and build2) were **not** updated to use the new parser, they are hardcoded to the old parser.
- These two methods are used by subclasses like URI::HTTP, for building URIs from parts: URI::HTTP.build(host: "foobar.com")
- The main significant difference is that the old parser fails with hostnames including underscores, which are now valid.

This issue was to fix these two methods, which are probably rarely used in comparison to URI() but are still useful.

- The issue has a patch attached.
- The patch resolves the issue, includes a test, and doesn't fail any other tests.
- The total diff is just +10 -6.

This is a very minor issue, which is probably why it's gone unnoticed for 8 years, but the fix is also very isolated and hopefully very low risk.

Thanks,
Gareth

#3 - 04/09/2024 07:34 PM - jeremyevans0 (Jeremy Evans)

I'm in favor of this change. However, be aware that uri is maintained at https://github.com/ruby/uri . Could you please submit a pull request to that repository?

#4 - 04/09/2024 07:43 PM - jeremyevans0 (Jeremy Evans)

- Related to Bug #19756: URI::HTTP.build does not accept a host of `_gateway`, but `URI.parse` will. added

#5 - 06/11/2024 01:11 AM - gareth (Gareth Adams)

jeremyevans0 (Jeremy Evans) wrote in #note-3:

> I'm in favor of this change. However, be aware that uri is maintained at https://github.com/ruby/uri . Could you please submit a pull request to that repository?

Thanks Jeremy, I've replicated this patch in https://github.com/ruby/uri/pull/105 if you're happy with the change :)

Gareth

#6 - 07/16/2024 02:18 AM - naruse (Yui NARUSE)

Since the use case sounds reasonable, let's try the new version.
@hsbt (Hiroshi SHIBATA) could you change it?

#7 - 07/16/2024 02:19 AM - hsbt (Hiroshi SHIBATA)

- Status changed from Open to Assigned

- Assignee set to hsbt (Hiroshi SHIBATA)

#8 - 07/18/2024 04:50 AM - hsbt (Hiroshi SHIBATA)

I implemented to use RFC3986 parser for URI library at https://github.com/ruby/uri/pull/107

It provides URI.parser= method for using RFC2396 parser after that.

```
URI.parser = URI::RFC2396_PARSER
```

The users can use RFC2396 parser if they need to URI::REGEXP module or previous behavior.

#9 - 07/22/2024 01:57 AM - hsbt (Hiroshi SHIBATA)

- Status changed from Assigned to Closed

I has been merged https://github.com/ruby/uri/pull/107

I consider to bump up version to 1.0.0 with the current URI HEAD. I update it to 1.0.0.pre1 for testing at next Ruby preview release like preview2.

Files

0001-Update-URI-Generic.build-build2-to-use-RFC3986_PARSE.patch	2.89 KB	01/10/2023	gareth (Gareth Adams)