
URL Documentation

Jon Cram

Dec 10, 2018

1	Overview	3
1.1	Url	3
1.2	Normalizer	3
1.3	Parser	4
2	Requirements and Installation	7
2.1	Requirements	7
2.2	Installation	7
3	URL Model	9
3.1	Creating a URL	9
3.2	Component Access	9
3.3	Component Modification	10
3.4	Non-Optional Normalization	11
4	Normalizer	13
4.1	Capitalize Percent Encoding	14
4.2	Decode Unreserved Characters	14
4.3	Convert Empty HTTP Path	14
4.4	Remove Default File Host	15
4.5	Remove Default Port	15
4.6	Remove Path Dot Segments	15
4.7	Convert Host Unicode to Punycode	16
4.8	Reduce Duplicate Path Slashes	16
4.9	Sort Query Parameters	16
4.10	Add Path Trailing Slash	17
4.11	Remove User Info	17
4.12	Remove Fragment	17
4.13	Remove www Sub-domain	17
4.14	Specify a Default Scheme	18
4.15	Remove Filenames From Path By Pattern	18
4.16	Remove Query Parameters By Pattern	18
5	Applying Multiple Normalizations	21
5.1	Multiple Flags	21
5.2	Options With No Other Normalizations	21
5.3	Apply All Semantically-Lossless Normalizations	22

6	Semantically-Lossless Normalizations	23
7	Potentially-Lossy Normalizations	25
8	Lossy Normalizations	27

The `webginition/url` package models, normalizes, parses and compares URLs.

The *overview* provides a high-level introduction to the package classes.

You might want to skip straight to the *Url usage guide* or the *Normalizer usage guide*.

1.1 Url

A `Url` models a [URL](#), providing component access and modification through the [PSR7 UriInterface](#).

```
<?php
use webignition\Url\Url;

$url = new Url('http://example.com/path?query#fragment');

$url->getScheme();
// "http"

$url->getQuery();
// "query"

$modifiedUrl = $url
    ->withScheme('https')
    ->withPath('/modified-path')
    ->withQuery('foo=bar')
    ->withFragment('');
(string) $modifiedUrl;
// https://example.com/modified-path?foo=bar
```

Read the *Url usage* guide for more detail.

1.2 Normalizer

The Normalizer can apply any combination of sixteen normalizations to any `UriInterface` implementation.

```
<?php
use webignition\Url\Normalizer;

$url = new Url('http://example.com/path?c=cow&a=apple&b=bear#fragment');

$normalizedUrl = Normalizer::normalize(
    $url,
    Normalizer::SORT_QUERY_PARAMETERS | Normalizer::REMOVE_FRAGMENT
);

(string) $normalizedUrl;
// "http://example.com/path?a=apple&b=bear&c=cow"
```

Flags:

- *capitalize percent encoding*
- *decode unreserved characters*
- *convert empty http path*
- *remove default file host*
- *remove port host*
- *remove path dot segments*
- *convert host unicode to punycode*
- *reduce duplicate path slashes*
- *sort query parameters*
- *add path trailing slash-*
- *remove user info*
- *remove fragment*
- *remove www sub-domain*

Options:

- *specify a default scheme*
- *remove filenames from path by pattern*
- *remove query parameters by pattern*

Read the *Normalizer usage* guide for more detail.

1.3 Parser

The `Parser` transforms a URL string into an array of component parts. Useful for direct access to raw URL components.

The parser is used internally by `Url::create()` and isn't needed for normal URL usage.

```
<?php
use webignition\Url\Parser;
```

(continues on next page)

(continued from previous page)

```
$components = Parser::parse('https://example.com:8080/path?query#fragment');  
  
$components[Parser::COMPONENT_SCHEME];  
// "https"  
  
$components[Parser::COMPONENT_HOST];  
// "example.com"
```

Requirements and Installation

2.1 Requirements

- PHP 7.2.0 or greater
- composer

2.2 Installation

```
composer require webignition/url ^3
```


A `Url` models a **URL**, providing component access and modification through the **PSR7 UriInterface**.

Minimal non-optional RFC 3986 normalization is applied by default:

- converts scheme to lowercase
- converts host to lowercase
- removes the default port

3.1 Creating a URL

A new `Url` instance is created by passing a URL string to the constructor:

```
<?php
use webignition\Url\Url;

$url = new Url('https://example.com');
```

3.2 Component Access

```
<?php
use webignition\Url\Url;

$url = new Url('https://user:password@example.com:8080/path?query#fragment');

$url->getScheme();
// "https"
```

(continues on next page)

```
$url->getUserInfo();  
// "user:password"  
  
$url->getHost();  
// "example.com"  
  
$url->getPort();  
// 8080  
  
$url->getAuthority();  
// "user:password@example.com:8080"  
  
$url->getPath();  
// "/path"  
  
$url->getQuery();  
// "query"  
  
$url->getFragment();  
// "fragment"
```

3.3 Component Modification

The `Url::with*()` are used to set components. A `Url` is immutable. The return value is a new `Url` instance.

```
<?php  
  
use webignition\Url\Url;  
  
$url = new Url('https://user:password@example.com:8080/path?query#fragment');  
(string) $url;  
// "https://user:password@example.com:8080/path?query#fragment"  
  
$url = $url->withScheme('http');  
(string) $modifiedUrl;  
// "http://user:password@example.com:8080/path?query#fragment"  
  
$url = $url->withUserInfo('new-user', 'new-password');  
(string) $modifiedUrl;  
// "http://new-user:new-password@example.com:8080/path?query#fragment"  
  
$url = $url->withUserInfo('');  
(string) $modifiedUrl;  
// "http://example.com:8080/path?query#fragment"  
  
$url = $url->withHost('new.example.com');  
(string) $modifiedUrl;  
// "http://new.example.com:8080/path?query#fragment"  
  
$url = $url->withPort(null);  
(string) $modifiedUrl;  
// "http://new.example.com/path?query#fragment"  
  
$url = $url->withPath('');
```

(continues on next page)

(continued from previous page)

```
(string) $modifiedUrl;  
// "http://new.example.com?query#fragment"  
  
$url = $url->withQuery('');  
(string) $modifiedUrl;  
// "http://new.example.com#fragment"  
  
$url = $url->withFragment('');  
(string) $modifiedUrl;  
// "http://new.example.com"
```

3.4 Non-Optional Normalization

```
<?php  
  
use webignition\Url\Url;  
  
$url = new Url('HTTPS://EXAMPLE.com:443');  
  
$url->getScheme();  
// "https"  
  
$url->getHost();  
// "example.com"  
  
$url->getPort();  
// null
```


The `Normalizer` can apply any combination of sixteen normalizations to any `UriInterface` implementation.

Normalizations are specified through either *flags* or *options*. Flags are for normalizations that can be turned on or turned off (you either want it or you don't). Options are for normalizations that act on one or more variables that you get to choose.

Flags:

- *capitalize percent encoding*
- *decode unreserved characters*
- *convert empty http path*
- *remove default file host*
- *remove port host*
- *remove path dot segments*
- *convert host unicode to punycode*
- *reduce duplicate path slashes*
- *sort query parameters*
- *add path trailing slash-*
- *remove user info*
- *remove fragment*
- *remove www sub-domain*

Options:

- *specify a default scheme*
- *remove filenames from path by pattern*
- *remove query parameters by pattern*

4.1 Capitalize Percent Encoding

Convert percent-encoded triplets (such as %3A) to uppercase. Letters within a percent-encoded triplet are case-insensitive.

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://example.com/path%2fvalue');
$normalizedUrl = Normalizer::normalize($url, Normalizer::CAPITALIZE_PERCENT_ENCODING);

(string) $normalizedUrl;
// "http://example.com/path%2Fvalue"
```

4.2 Decode Unreserved Characters

Convert percent-encoded characters that have no special meaning to their unencoded equivalents.

Decodes encoded forms of: ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789-._~

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://example.com/%75%72%6C');
$normalizedUrl = Normalizer::normalize($url, Normalizer::DECODE_UNRESERVED_
↳CHARACTERS);

(string) $normalizedUrl;
// "http://example.com/url"
```

4.3 Convert Empty HTTP Path

Applies a path of / where the path is empty and the scheme is http or https.

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://example.com');
$normalizedUrl = Normalizer::normalize($url, Normalizer::CONVERT_EMPTY_HTTP_PATH);

(string) $normalizedUrl;
// "http://example.com/"
```

4.4 Remove Default File Host

Removes the host of `localhost` from a `file://` url.

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('file://localhost/path');
$normalizedUrl = Normalizer::normalize($url, Normalizer::REMOVE_DEFAULT_FILE_HOST);

(string) $normalizedUrl;
// "file:///path"
```

4.5 Remove Default Port

Removes the port if it matches the default port for the scheme.

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://example.com:80');
$normalizedUrl = Normalizer::normalize($url, Normalizer::REMOVE_DEFAULT_PORT);

(string) $normalizedUrl;
// "http://example.com"

$url = new Url('https://example.com:443');
$normalizedUrl = Normalizer::normalize($url, Normalizer::REMOVE_DEFAULT_PORT);

(string) $normalizedUrl;
// "https://example.com"
```

4.6 Remove Path Dot Segments

The `.` and `..` path segments have a special meaning. These segments are removed and the path is re-written to be equivalent.

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://example.com/a/b/c/../../g');
$normalizedUrl = Normalizer::normalize($url, Normalizer::REMOVE_PATH_DOT_SEGMENTS);

(string) $normalizedUrl;
// "http://example.com/a/g"
```

4.7 Convert Host Unicode to Punycode

Unicode hosts containing non-ascii characters are converted to the `punycode` equivalent.

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://.example.com/');
$normalizedUrl = Normalizer::normalize($url, Normalizer::CONVERT_HOST_UNICODE_TO_
    ↳PUNYCODE);

(string) $normalizedUrl;
// "http://xn--g6h.example.com/"
```

4.8 Reduce Duplicate Path Slashes

Reduces occurrences of multiple slashes in the path to single slashes.

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://example.com///path//');
$normalizedUrl = Normalizer::normalize($url, Normalizer::REDUCE_DUPLICATE_PATH_
    ↳SLASHES);

(string) $normalizedUrl;
// "http://example.com/path/"
```

4.9 Sort Query Parameters

Alphabetically sorts query parameters by key.

Sorting is neither locale- nor unicode-aware. The purpose is to be able to compare URLs in a reproducible way.

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://example.com?c=cow&a=apple&b=bear');
$normalizedUrl = Normalizer::normalize($url, Normalizer::SORT_QUERY_PARAMETERS);

(string) $normalizedUrl;
// "http://example.com?a=apple&b=bear&c=cow"
```

4.10 Add Path Trailing Slash

Add a trailing slash to the path if not present.

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://example.com');
$normalizedUrl = Normalizer::normalize($url, Normalizer::ADD_PATH_TRAILING_SLASH);

(string) $normalizedUrl;
// "http://example.com/"
```

4.11 Remove User Info

Remove user credentials.

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://user:password@example.com');
$normalizedUrl = Normalizer::normalize($url, Normalizer::REMOVE_USER_INFO);

(string) $normalizedUrl;
// "http://example.com"
```

4.12 Remove Fragment

Remove fragment component.

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://example.com#fragment');
$normalizedUrl = Normalizer::normalize($url, Normalizer::REMOVE_FRAGMENT);

(string) $normalizedUrl;
// "http://example.com"
```

4.13 Remove www Sub-domain

Remove the www sub-domain.

```
<?php

use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://www.example.com');
$normalizedUrl = Normalizer::normalize($url, Normalizer::REMOVE_WWW);

(string) $normalizedUrl;
// "http://example.com"
```

4.14 Specify a Default Scheme

Specify a default scheme to be applied if none is present.

```
<?php

use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('//www.example.com');
$normalizedUrl = Normalizer::normalize($url, Normalizer::NONE, [
    Normalizer::OPTION_DEFAULT_SCHEME => 'http',
]);

(string) $normalizedUrl;
// "http://example.com"
```

4.15 Remove Filenames From Path By Pattern

Remove the filename from the path component. Removal is defined through one or more patterns.

Useful for stripping common default filenames such as `index.html`, `index.js` or `default.asp`.

```
<?php

use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://www.example.com/index.html');
$normalizedUrl = Normalizer::normalize($url, Normalizer::NONE, [
    Normalizer::OPTION_REMOVE_PATH_FILES_PATTERNS => Normalizer::REMOVE_INDEX_FILE_
    ↪PATTERN,
]);

(string) $normalizedUrl;
// "http://example.com/"
```

4.16 Remove Query Parameters By Pattern

Remove query parameters where the parameter key matches one of a set of patterns.

Useful for stripping query parameters considered by you to be irrelevant to the canonical form of a URL.

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://www.example.com?x=1&y=2&utm_source=facebook&utm_medium=18');
$normalizedUrl = Normalizer::normalize($url, Normalizer::NONE, [
    Normalizer::OPTION_REMOVE_QUERY_PARAMETERS_PATTERNS => [
        '/^utm_\w+/i',
    ],
]);

(string) $normalizedUrl;
// "http://example.com?page=1&category=2"
```

Applying Multiple Normalizations

You can apply any number of the sixteen normalizations when normalizing a `UriInterface` instance.

5.1 Multiple Flags

Combine flags using the bitwise `|` operator.

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://example.com?b=bear&a=apple#fragment');
$normalizedUrl = Normalizer::normalize(
    $url,
    Normalizer::SORT_QUERY_PARAMETERS | Normalizer::REMOVE_FRAGMENT
);

(string) $normalizedUrl;
// "http://example.com?a=apple&b=bear"
```

5.2 Options With No Other Normalizations

To apply one or more options but no other normalizations, call `Normalizer::normalize()` with `Normalizer::NONE` as the flags argument.

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;
```

(continues on next page)

(continued from previous page)

```
$url = new Url('http://www.example.com?x=1&y=2&utm_source=facebook&utm_medium=18');
$normalizedUrl = Normalizer::normalize($url, Normalizer::NONE, [
    Normalizer::OPTION_REMOVE_QUERY_PARAMETERS_PATTERNS => [
        '/^utm_\w+/i',
    ],
]);

(string) $normalizedUrl;
// "http://example.com?page=1&category=2"
```

5.3 Apply All Semantically-Lossless Normalizations

A set of normalizations that do not change the semantics of a URL are defined as `Normalizer::PRESERVING_NORMALIZATIONS`.

Read more about *semantically-lossless normalizations* to see what flags this applies.

```
<?php

use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://.example.com:80/p%61th/..?option=%3f');
$normalizedUrl = Normalizer::normalize($url, Normalizer::PRESERVING_NORMALIZATIONS);

(string) $normalizedUrl;
// "http://xn--g6h.example.com:80/path/?option=%3F"
```

The `flags` argument of `Normalizer::normalize()` defaults to `Normalizer::PRESERVING_NORMALIZATIONS`.

The following is equivalent to the above:

```
<?php

use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://.example.com:80/p%61th/..?option=%3f');
$normalizedUrl = Normalizer::normalize($url);

(string) $normalizedUrl;
// "http://xn--g6h.example.com:80/path/?option=%3F"
```

Semantically-Lossless Normalizations

There is a set of normalizations that do not change the semantics of a URL. These are defined as `Normalizer::PRESERVING_NORMALIZATIONS`. The normalizer applies this set of normalizations if no specific normalizations are requested.

- *capitalize percent encoding*
- *decode unreserved characters*
- *convert empty http path*
- *remove default file host*
- *remove port host*
- *remove path dot segments*
- *convert host unicode to punycode*

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;

$url = new Url('http://.example.com:80/p%61th/..?option=%3f');
$normalizedUrl = Normalizer::normalize($url);

(string) $normalizedUrl;
// "http://xn--g6h.example.com:80/path/?option=%3F"
```

The `Normalizer::PRESERVING_NORMALIZATIONS` flag can be used in conjunction with additional normalizations.

```
<?php
use webignition\Url\Normalizer;
use webignition\Url\Url;
```

(continues on next page)

(continued from previous page)

```
$url = new Url('http://.example.com:80/p%61th/../?option=%3f&b=bear&a-apple');
$normalizedUrl = Normalizer::normalize(
    $url,
    Normalizer::PRESERVING_NORMALIZATIONS |
    Normalizer::SORT_QUERY_PARAMETERS
);

(string) $normalizedUrl;
// "http://xn--g6h.example.com:80/path/?a=apple&bear&option=%3F"
```

Potentially-Lossy Normalizations

The reduce path slashes and sort query parameters normalizations are potentially lossy. Semantic equivalence may be commonly achieved but is not guaranteed.

- *reduce duplicate path slashes*
- *sort query parameters*

Lossy Normalizations

Lossy normalizations change, to a greater or lesser extent, the semantics of a URL.

- *add path trailing slash-*
- *remove user info*
- *remove fragment*
- *remove www sub-domain*