

Package: BCP47 (via r-universe)

June 8, 2026

Title Work with Language Tags

Version 0.0.0.9000

Description Tools to parse, validate, normalize, and match language tags following the Best Current Practice 47 (BCP 47) standard, which defines the syntax (RFC 5646, <<https://tools.ietf.org/html/rfc5646>>) and lookup rules (RFC 4647, <<https://tools.ietf.org/html/rfc4647>>) for identifying human languages. Includes a bundled snapshot of the IANA Language Subtag Registry (<<https://www.iana.org/assignments/language-subtag-registry/>>) with optional support for updating.

License MIT + file LICENSE

Depends R (>= 4.2.0)

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.3

Imports purrr, tibble

URL <https://christophertkenny.com/BCP47/>

Suggests knitr, rmarkdown, testthat (>= 3.0.0)

Config/testthat/edition 3

Config/Needs/website christopherkenny/ctktemplate

VignetteBuilder knitr

Repository <https://christopherkenny.r-universe.dev>

Date/Publication 2026-05-01 16:14:47 UTC

RemoteUrl <https://github.com/christopherkenny/BCP47>

RemoteRef HEAD

RemoteSha 7d680fdd7cab5a4a7f1732c75ad127a0066049cf

Contents

bcp_cache_path	2
bcp_match_language	3
bcp_normalize	3
bcp_parse	4
bcp_process_registry	5
bcp_validate	6

Index	7
--------------	----------

bcp_cache_path	<i>Work with the BCP47 cache</i>
----------------	----------------------------------

Description

bcp_cache_path() returns the path to the cached registry file. bcp_cache_size() reports the total size of cache files on disk. bcp_cache_clear() deletes the cached registry file. bcp_cache_update() downloads the latest IANA registry and writes it to cache.

Usage

```
bcp_cache_path()

bcp_cache_size()

bcp_cache_clear(force = FALSE)

bcp_cache_update(overwrite = FALSE)
```

Arguments

force	Logical. If TRUE, skip interactive confirmation in bcp_cache_clear(). Default is FALSE.
overwrite	Logical. If TRUE, overwrite an existing cached registry in bcp_cache_update(). Default is FALSE.

Value

- bcp_cache_path(): A character scalar giving the file path of the cached registry.
- bcp_cache_size(): The cache size in bytes, invisibly. Also prints a human-readable size message.
- bcp_cache_clear(): The cache path, invisibly.
- bcp_cache_update(): The registry data frame, invisibly.

Examples

```
bcp_cache_path()
bcp_cache_size()
```

bcp_match_language	<i>Match a BCP 47 language preference to available languages</i>
--------------------	------------------------------------------------------------------

Description

Implements the "Lookup" scheme from RFC 4647 Section 3.4: given an ordered list of language tag preferences and a set of available tags, returns the best match. Matching proceeds by progressively stripping the rightmost subtag from each preference until a match is found or no subtags remain. Single-character subtags (including the private-use prefix "x") are always stripped as a unit together with any preceding single-character subtags.

Usage

```
bcp_match_language(preferences, available, default = NULL)
```

Arguments

preferences	A character vector of BCP 47 language tags representing the caller's ordered preferences, from most to least preferred.
available	A character vector of BCP 47 language tags that are available to choose from.
default	The value to return when no preference matches any available tag. Defaults to NULL.

Value

The best-matching tag from available, or default if no match is found. Matching is case-insensitive; the returned value preserves the original casing from available.

Examples

```
bcp_match_language(c('en-US', 'fr'), c('en', 'fr-FR', 'de'))  
bcp_match_language('zh-Hans-CN', c('zh-TW', 'zh-Hans', 'en'))  
bcp_match_language('pt-BR', c('fr', 'de'), default = 'en')
```

bcp_normalize	<i>Normalize a BCP 47 language tag</i>
---------------	----------------------------------------

Description

Applies the canonicalization rules from RFC 5646: preferred values are substituted for deprecated subtags, default scripts are suppressed, and each component is cased according to convention (language lower-case, script title-case, region upper-case).

Usage

```
bcp_normalize(tag, registry = bcp_get_registry())
```

Arguments

tag	A character scalar BCP 47 language tag.
registry	A data frame of the IANA Language Subtag Registry, as returned by <code>bcp_process_registry()</code> . Defaults to the locally cached registry, falling back to the bundled snapshot if no cache exists.

Value

A character scalar with the normalized BCP 47 tag.

Examples

```
bcp_normalize('en-us')
bcp_normalize('ZH-hans-cn')
```

bcp_parse	<i>Parse a BCP 47 language tag</i>
-----------	------------------------------------

Description

Decomposes a BCP 47 language tag into its constituent subtags following the syntax defined in RFC 5646. Both hyphen (-) and underscore (_) are accepted as subtag separators.

Usage

```
bcp_parse(tag)
```

Arguments

tag	A character scalar BCP 47 language tag.
-----	-----------------------------------------

Value

A named list with the following elements:

language	The primary language subtag (e.g., "en", "zh"), or NA for a pure private-use tag.
extlang	A character vector of extended language subtags (three-letter codes following the primary language), or NULL.
script	The four-letter script subtag (e.g., "latn", "hans"), or NA if absent.
region	The two-letter or three-digit region subtag (e.g., "us", "419"), or NA if absent.
variants	A character vector of variant subtags, or NULL.
extensions	A named list of extension subtag sequences, keyed by the single-letter extension singleton.
private	A character vector of private-use subtags (following x-), or NULL.

All subtags are returned in lower-case.

Examples

```
bcp_parse('en-US')
bcp_parse('zh-Hans-CN')
bcp_parse('de-1901')
bcp_parse('x-private')
```

bcp_process_registry *Process the IANA BCP 47 subtag registry*

Description

Downloads and parses the IANA Language Subtag Registry into a tidy data frame. Each row represents one registry entry (language, extlang, script, region, variant, grandfathered, or redundant tag). Columns correspond to registry fields such as type, subtag, description, added, preferred_value, and suppress_script. When an entry has multiple values for a single field (e.g., multiple Description lines), they are joined with ";".

Usage

```
bcp_process_registry(
  url =
    "https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry"
)
```

Arguments

url A character scalar giving the URL of the registry plain-text file. Defaults to the official IANA location.

Value

A tibble with one row per registry entry and one column per field. The last_update attribute records the File-Date from the registry header.

Examples

```
reg <- bcp_process_registry()
head(reg)
```

bcp_validate	<i>Validate a BCP 47 language tag</i>
--------------	---------------------------------------

Description

Checks whether the primary language, script, and region subtags of a BCP 47 tag appear in the IANA Language Subtag Registry. Extension and private-use subtags are not validated.

Usage

```
bcp_validate(tag, registry = bcp_get_registry())
```

Arguments

tag	A character scalar BCP 47 language tag.
registry	A data frame of the IANA Language Subtag Registry, as returned by bcp_process_registry() . Defaults to the locally cached registry, falling back to the bundled snapshot if no cache exists.

Value

A logical scalar: TRUE if all checked subtags are registered, FALSE otherwise.

Examples

```
bcp_validate('en-US')  
bcp_validate('en-Latn-US')  
bcp_validate('xx-ZZ')
```

Index

`bcp_cache_clear (bcp_cache_path)`, 2
`bcp_cache_path`, 2
`bcp_cache_size (bcp_cache_path)`, 2
`bcp_cache_update (bcp_cache_path)`, 2
`bcp_match_language`, 3
`bcp_normalize`, 3
`bcp_parse`, 4
`bcp_process_registry`, 5
`bcp_process_registry()`, 4, 6
`bcp_validate`, 6