Libidn2

Internationalized Domain Names (IDNA2008)
Version 0.10, 25 June 2014

Simon Josefsson

This manual is for Libidn2 (version 0.10, 25 June 2014), an implementation of IDNA2008 internationalized domain names.

Table of Contents

1 Introduction

Libidn2 is a free software implementation of IDNA2008.

2 Library Functions

Below are the interfaces of the Libidn2 library documented.

2.1 Header file `idn2.h`

To use the functions documented in this chapter, you need to include the file `idn2.h` like this:

```
#include <idn2.h>
```

2.2 Core Functions

When you have the data encoded in UTF-8 form the direct interfaces to the library are as follows.

idn2_lookup_u8

`int idn2_lookup_u8` (*const uint8_t * **src**, uint8_t ** **lookupname**, int **flags***) [Function]

src: input zero-terminated UTF-8 string in Unicode NFC normalized form.

lookupname: newly allocated output variable with name to lookup in DNS.

flags: optional `idn2_flags` to modify behaviour.

Perform IDNA2008 lookup string conversion on domain name `src`, as described in section 5 of RFC 5891. Note that the input string must be encoded in UTF-8 and be in Unicode NFC form.

Pass `IDN2_NFC_INPUT` in `flags` to convert input to NFC form before further processing. Pass `IDN2_ALABEL_ROUNDTRIP` in `flags` to convert any input A-labels to U-labels and perform additional testing. Multiple flags may be specified by binary or:ing them together, for example `IDN2_NFC_INPUT | IDN2_ALABEL_ROUNDTRIP`.

Returns: On successful conversion `IDN2_OK` is returned, if the output domain or any label would have been too long `IDN2_TOO_BIG_DOMAIN` or `IDN2_TOO_BIG_LABEL` is returned, or another error code is returned.

idn2_register_u8

`int idn2_register_u8` (*const uint8_t * **ulabel**, const uint8_t * **alabel**, uint8_t ** **insertname**, int **flags***) [Function]

ulabel: input zero-terminated UTF-8 and Unicode NFC string, or NULL.

alabel: input zero-terminated ACE encoded string (xn–), or NULL.

insertname: newly allocated output variable with name to register in DNS.

flags: optional `idn2_flags` to modify behaviour.

Perform IDNA2008 register string conversion on domain label `ulabel` and `alabel`, as described in section 4 of RFC 5891. Note that the input `ulabel` must be encoded in UTF-8 and be in Unicode NFC form.

Pass `IDN2_NFC_INPUT` in `flags` to convert input `ulabel` to NFC form before further processing.

It is recommended to supply both `ulabel` and `alabel` for better error checking, but supplying just one of them will work. Passing in only `alabel` is better than only `ulabel`. See RFC 5891 section 4 for more information.

Returns: On successful conversion `IDN2_OK` is returned, when the given `ulabel` and `alabel` does not match each other `IDN2_UALABEL_MISMATCH` is returned, when either of the input labels are too long `IDN2_TOO_BIG_LABEL` is returned, when `alabel` does does not appear to be a proper A-label `IDN2_INVALID_ALABEL` is returned, or another error code is returned.

2.3 Locale Functions

As a convenience, the following functions are provided that will convert the input from the locale encoding format to UTF-8 and normalize the string using NFC, and then apply the core functions described earlier.

idn2_lookup_ul

`int idn2_lookup_ul` (*const char * **src**, char ** **lookupname**, int **flags***) [Function]

src: input zero-terminated locale encoded string.

lookupname: newly allocated output variable with name to lookup in DNS.

flags: optional `idn2_flags` to modify behaviour.

Perform IDNA2008 lookup string conversion on domain name `src`, as described in section 5 of RFC 5891. Note that the input is assumed to be encoded in the locale's default coding system, and will be transcoded to UTF-8 and NFC normalized by this function.

Pass `IDN2_ALABEL_ROUNDTRIP` in `flags` to convert any input A-labels to U-labels and perform additional testing.

Returns: On successful conversion `IDN2_OK` is returned, if conversion from locale to UTF-8 fails then `IDN2_ICONV_FAIL` is returned, if the output domain or any label would have been too long `IDN2_TOO_BIG_DOMAIN` or `IDN2_TOO_BIG_LABEL` is returned, or another error code is returned.

idn2_register_ul

`int idn2_register_ul` (*const char * **ulabel**, const char * **alabel**, char ** **insertname**, int **flags***) [Function]

ulabel: input zero-terminated locale encoded string, or NULL.

alabel: input zero-terminated ACE encoded string (xn–), or NULL.

insertname: newly allocated output variable with name to register in DNS.

flags: optional `idn2_flags` to modify behaviour.

Perform IDNA2008 register string conversion on domain label `ulabel` and `alabel`, as described in section 4 of RFC 5891. Note that the input `ulabel` is assumed to be encoded in the locale's default coding system, and will be transcoded to UTF-8 and NFC normalized by this function.

It is recommended to supply both `ulabel` and `alabel` for better error checking, but supplying just one of them will work. Passing in only `alabel` is better than only `ulabel`. See RFC 5891 section 4 for more information.

Returns: On successful conversion `IDN2_OK` is returned, when the given `ulabel` and `alabel` does not match each other `IDN2_UALABEL_MISMATCH` is returned, when either of the input labels are too long `IDN2_TOO_BIG_LABEL` is returned, when `alabel` does does not appear to be a proper A-label `IDN2_INVALID_ALABEL` is returned, or another error code is returned.

2.4 Control Flags

The `flags` parameter can take on the following values, or a bit-wise inclusive or of any subset of the parameters:

`idn2_flags IDN2_NFC_INPUT` [Global flag]
Apply NFC normalization on input.

`idn2_flags IDN2_ALABEL_ROUNDTRIP` [Global flag]
Apply additional round-trip conversion of A-label inputs.

2.5 Error Handling

idn2_strerror

`const char * idn2_strerror` (*int* `rc`) [Function]
rc: return code from another libidn2 function.

Convert internal libidn2 error code to a humanly readable string. The returned pointer must not be de-allocated by the caller.

Return value: A humanly readable string describing error.

idn2_strerror_name

`const char * idn2_strerror_name` (*int* `rc`) [Function]
rc: return code from another libidn2 function.

Convert internal libidn2 error code to a string corresponding to internal header file symbols. For example, idn2_strerror_name(IDN2_MALLOC) will return the string "IDN2_MALLOC".

The caller must not attempt to de-allocate the returned string.

Return value: A string corresponding to error code symbol.

2.6 Return Codes

The functions normally return 0 on sucess or a negative error code.

`idn2_rc IDN2_OK` [Return code]
Successful return.

`idn2_rc IDN2_MALLOC` [Return code]
Memory allocation error.

`idn2_rc IDN2_NO_CODESET` [Return code]
: Could not determine locale string encoding format.

`idn2_rc IDN2_ICONV_FAIL` [Return code]
: Could not transcode locale string to UTF-8.

`idn2_rc IDN2_ENCODING_ERROR` [Return code]
: Unicode data encoding error.

`idn2_rc IDN2_NFC` [Return code]
: Error normalizing string.

`idn2_rc IDN2_PUNYCODE_BAD_INPUT` [Return code]
: Punycode invalid input.

`idn2_rc IDN2_PUNYCODE_BIG_OUTPUT` [Return code]
: Punycode output buffer too small.

`idn2_rc IDN2_PUNYCODE_OVERFLOW` [Return code]
: Punycode conversion would overflow.

`idn2_rc IDN2_TOO_BIG_DOMAIN` [Return code]
: Domain name longer than 255 characters.

`idn2_rc IDN2_TOO_BIG_LABEL` [Return code]
: Domain label longer than 63 characters.

`idn2_rc IDN2_INVALID_ALABEL` [Return code]
: Input A-label is not valid.

`idn2_rc IDN2_UALABEL_MISMATCH` [Return code]
: Input A-label and U-label does not match.

`idn2_rc IDN2_NOT_NFC` [Return code]
: String is not NFC.

`idn2_rc IDN2_2HYPHEN` [Return code]
: String has forbidden two hyphens.

`idn2_rc IDN2_HYPHEN_STARTEND` [Return code]
: String has forbidden starting/ending hyphen.

`idn2_rc IDN2_LEADING_COMBINING` [Return code]
: String has forbidden leading combining character.

`idn2_rc IDN2_DISALLOWED` [Return code]
: String has disallowed character.

`idn2_rc IDN2_CONTEXTJ` [Return code]
: String has forbidden context-j character.

`idn2_rc IDN2_CONTEXTJ_NO_RULE` [Return code]
: String has context-j character with no rull.

`idn2_rc IDN2_CONTEXTO` [Return code]
String has forbidden context-o character.

`idn2_rc IDN2_CONTEXTO_NO_RULE` [Return code]
String has context-o character with no rull.

`idn2_rc IDN2_UNASSIGNED` [Return code]
String has forbidden unassigned character.

`idn2_rc IDN2_BIDI` [Return code]
String has forbidden bi-directional properties.

2.7 Memory Handling

idn2_free

`void idn2_free (`*`void * ptr`*`)` [Function]
ptr: pointer to deallocate

Call free(3) on the given pointer.

This function is typically only useful on systems where the library malloc heap is different from the library caller malloc heap, which happens on Windows when the library is a separate DLL.

2.8 Version Check

It is often desirable to check that the version of Libidn2 used is indeed one which fits all requirements. Even with binary compatibility new features may have been introduced but due to problem with the dynamic linker an old version is actually used. So you may want to check that the version is okay right after program startup.

idn2_check_version

`const char * idn2_check_version (`*`const char * req_version`*`)` [Function]
req_version: version string to compare with, or NULL.

Check IDN2 library version. This function can also be used to read out the version of the library code used. See `IDN2_VERSION` for a suitable `req_version` string, it corresponds to the idn2.h header file version. Normally these two version numbers match, but if you are using an application built against an older libidn2 with a newer libidn2 shared library they will be different.

Return value: Check that the version of the library is at minimum the one given as a string in `req_version` and return the actual version string of the library; return NULL if the condition is not met. If NULL is passed to this function no check is done and only the version string is returned.

The normal way to use the function is to put something similar to the following first in your `main`:

```
if (!idn2_check_version (IDN2_VERSION))
  {
```

```
  printf ("idn2_check_version() failed:\n"
          "Header file incompatible with shared library.\n");
  exit(EXIT_FAILURE);
}
```

3 Examples

This chapter contains example code which illustrate how Libidn2 is used when you write your own application.

3.1 Lookup

This example demonstrates how a domain name is processed before it is lookup in the DNS.

```
#include <stdio.h> /* printf, fflush, fgets, stdin, perror, fprintf */
#include <string.h> /* strlen */
#include <locale.h> /* setlocale */
#include <stdlib.h> /* free */
#include <idn2.h> /* idn2_lookup_ul, IDN2_OK, idn2_strerror, idn2_strerror_name */

int
main (int argc, char *argv[])
{
  int rc;
  char src[BUFSIZ];
  char *lookupname;

  setlocale (LC_ALL, "");

  printf ("Enter (possibly non-ASCII) domain name to lookup:  ");
  fflush (stdout);
  if (!fgets (src, sizeof (src), stdin))
    {
      perror ("fgets");
      return 1;
    }
  src[strlen (src) - 1] = '\0';

  rc = idn2_lookup_ul (src, &lookupname, 0);
  if (rc != IDN2_OK)
    {
      fprintf (stderr, "error:  %s (%s, %d)\n",
               idn2_strerror (rc), idn2_strerror_name (rc), rc);
      return 1;
    }

  printf ("IDNA2008 domain name to lookup in DNS: %s\n", lookupname);

  free (lookupname);

  return 0;
}
```

3.2 Register

This example demonstrates how a domain label is processed before it is registered in the DNS.

```
#include <stdio.h> /* printf, fflush, fgets, stdin, perror, fprintf */
#include <string.h> /* strlen */
#include <locale.h> /* setlocale */
#include <stdlib.h> /* free */
#include <idn2.h> /* idn2_register_ul, IDN2_OK, idn2_strerror, idn2_strerror_name */

int
main (int argc, char *argv[])
{
  int rc;
  char src[BUFSIZ];
  char *insertname;

  setlocale (LC_ALL, "");

  printf ("Enter (possibly non-ASCII) label to register:  ");
  fflush (stdout);
  if (!fgets (src, sizeof (src), stdin))
    {
      perror ("fgets");
      return 1;
    }
  src[strlen (src) - 1] = '\0';

  rc = idn2_register_ul (src, NULL, &insertname, 0);
  if (rc != IDN2_OK)
    {
      fprintf (stderr, "error:  %s (%s, %d)\n",
               idn2_strerror (rc), idn2_strerror_name (rc), rc);
      return 1;
    }

  printf ("IDNA2008 label to register in DNS: %s\n", insertname);

  free (insertname);

  return 0;
}
```

4 Invoking idn2

`idn2` translates internationalized domain names to the IDNA2008 encoded format, either for lookup or registration.

If strings are specified on the command line, they are used as input and the computed output is printed to standard output `stdout`. If no strings are specified on the command line, the program read data, line by line, from the standard input `stdin`, and print the computed output to standard output. What processing is performed (e.g., lookup or register) is indicated by options. If any errors are encountered, the execution of the applications is aborted.

All strings are expected to be encoded in the preferred charset used by your locale. Use `--debug` to find out what this charset is. On POSIX systems you may use the `LANG` environment variable to specify a different locale.

To process a string that starts with `-`, for example `-foo`, use `--` to signal the end of parameters, as in `idn2 -r -- -foo`.

4.1 Options

`idn2` recognizes these commands:

```
-h, --help              Print help and exit

-V, --version           Print version and exit

-l, --lookup            Lookup domain name (default)

-r, --register          Register label

    --debug             Print debugging information

    --quiet             Silent operation
```

4.2 Environment Variables

On POSIX systems the *LANG* environment variable can be used to override the system locale for the command being invoked. The system locale may influence what character set is used to decode data (i.e., strings on the command line or data read from the standard input stream), and to encode data to the standard output. If your system is set up correctly, however, the application will use the correct locale and character set automatically. Example usage:

```
$ LANG=en_US.UTF-8 idn2
...
```

4.3 Examples

Standard usage, reading input from standard input and disabling license and usage instructions:

```
jas@latte:~$ idn2 --quiet
räksmörgås.se
xn--rksmrgs-5wao1o.se
...
```

Reading input from the command line:

```
jas@latte:~$ idn2 räksmörgås.se blåbærgrød.no
xn--rksmrgs-5wao1o.se
xn--blbrgrd-fxak7p.no
jas@latte:~$
```

Testing the IDNA2008 Register function:

```
jas@latte:~$ idn2 --register fußball
xn--fuball-cta
jas@latte:~$
```

4.4 Troubleshooting

Getting character data encoded right, and making sure Libidn2 use the same encoding, can be difficult. The reason for this is that most systems may encode character data in more than one character encoding, i.e., using `UTF-8` together with `ISO-8859-1` or `ISO-2022-JP`. This problem is likely to continue to exist until only one character encoding come out as the evolutionary winner, or (more likely, at least to some extents) forever.

The first step to troubleshooting character encoding problems with Libidn2 is to use the '`--debug`' parameter to find out which character set encoding '`idn2`' believe your locale uses.

```
jas@latte:~$ idn2 --debug --quiet ""
Charset: UTF-8

jas@latte:~$
```

If it prints `ANSI_X3.4-1968` (i.e., `US-ASCII`), this indicate you have not configured your locale properly. To configure the locale, you can, for example, use '`LANG=sv_SE.UTF-8; export LANG`' at a `/bin/sh` prompt, to set up your locale for a Swedish environment using `UTF-8` as the encoding.

Sometimes '`idn2`' appear to be unable to translate from your system locale into `UTF-8` (which is used internally), and you will get an error message like this:

```
idn2: lookup: could not convert string to UTF-8
```

One explanation is that you didn't install the '`iconv`' conversion tools. You can find it as a standalone library in GNU Libiconv (http://www.gnu.org/software/libiconv/). On many GNU/Linux systems, this library is part of the system, but you may have to install additional packages to be able to use it.

Another explanation is that the error is correct and you are feeding '`idn2`' invalid data. This can happen inadvertently if you are not careful with the character set encoding you use. For example, if your shell run in a `ISO-8859-1` environment, and you invoke '`idn2`' with the '`LANG`' environment variable as follows, you will feed it `ISO-8859-1` characters but force it to believe they are `UTF-8`. Naturally this will lead to an error, unless the byte sequences happen to be valid `UTF-8`. Note that even if you don't get an error, the output may be

incorrect in this situation, because `ISO-8859-1` and `UTF-8` does not in general encode the same characters as the same byte sequences.

```
jas@latte:~$ idn2 --quiet --debug ""
Charset: ISO-8859-1

jas@latte:~$ LANG=sv_SE.UTF-8 idn2 --debug räksmörgås
Charset: UTF-8
input[0] = 0x72
input[1] = 0xc3
input[2] = 0xa4
input[3] = 0xc3
input[4] = 0xa4
input[5] = 0x6b
input[6] = 0x73
input[7] = 0x6d
input[8] = 0xc3
input[9] = 0xb6
input[10] = 0x72
input[11] = 0x67
input[12] = 0xc3
input[13] = 0xa5
input[14] = 0x73
UCS-4 input[0] = U+0072
UCS-4 input[1] = U+00e4
UCS-4 input[2] = U+00e4
UCS-4 input[3] = U+006b
UCS-4 input[4] = U+0073
UCS-4 input[5] = U+006d
UCS-4 input[6] = U+00f6
UCS-4 input[7] = U+0072
UCS-4 input[8] = U+0067
UCS-4 input[9] = U+00e5
UCS-4 input[10] = U+0073
output[0] = 0x72
output[1] = 0xc3
output[2] = 0xa4
output[3] = 0xc3
output[4] = 0xa4
output[5] = 0x6b
output[6] = 0x73
output[7] = 0x6d
output[8] = 0xc3
output[9] = 0xb6
output[10] = 0x72
output[11] = 0x67
output[12] = 0xc3
```

```
output[13] = 0xa5
output[14] = 0x73
UCS-4 output[0] = U+0072
UCS-4 output[1] = U+00e4
UCS-4 output[2] = U+00e4
UCS-4 output[3] = U+006b
UCS-4 output[4] = U+0073
UCS-4 output[5] = U+006d
UCS-4 output[6] = U+00f6
UCS-4 output[7] = U+0072
UCS-4 output[8] = U+0067
UCS-4 output[9] = U+00e5
UCS-4 output[10] = U+0073
xn--rksmrgs-5waap8p
jas@latte:~$
```

The sense moral here is to forget about '`LANG`' (instead, configure your system locale properly) unless you know what you are doing, and if you want to use '`LANG`', do it carefully and after verifying with '`--debug`' that you get the desired results.

Interface Index

Concept Index