Reviewing Programming Languages and Frameworks for Compliance With Universal Acceptance

Test Data Sets

Table of contents

Preamble

As noted in the main document, the sample test data given in the document at present is illustrative and deliberately incomplete. We recognise that practical experience of performing evaluations is necessary to guide the development of test data that will be of practical benefit in delivering useful evaluation results. This separate addendum document gives the design outline for test data sets for Universal Acceptance programming library assessment.

As noted, sharing of the test data sets developed will be highly beneficial to this process.

Requirements for test data sets

These tests are not designed to perform an exhaustive probe for all possible implementation problems in a library. Rather, the aim of the tests is to provide

- an indicator of the usefulness of the library;
- is the implementation generally correct, likely to be useful for most practical applications?
- are the common and some corner error cases correctly detected?

To that end, the tests are divided into *General* and *Specific* tests.

- *General* tests are tests that check overall common functionality works as expected; each test should be run several times with inputs reflecting cases likely to reflect real-world use, and each run counted as a different test result. To given a more concrete example, *general* domain name tests for Unicode base multilingual plane support should be run with input domain names from a variety of common scripts that might approximate expected real-world use, e.g. French, Chinese, Hindi, and Arabic. In this example, a single test will produce 4 test results.

- *Specific* tests check one particular piece of functionality works to specification; they have only the minimum input required to give an answer to the question posed by the test case. So, for example, a single instance of a domain with a Unicode combining mark as the first character is sufficient for test L-U2AS6 below. Similarly, some *specific* tests may in fact be covered by data from a *general* test; for example, test L-U2AS1 is likely to be covered by data for L-U2AG5. These test descriptions are included for completeness, but a test should be omitted, as it would be merely repeating an existing test.

The goal here is to ensure that the overall ratio of the tests passed to tests failed reflects the general usefulness and quality of the library.

Test Data Sets

The following sections specify the different tests for each category. Each test includes a reference to the standards document that specifies the behaviour the test is examining. There are several relevant standards, and they are not always completely consistent; hence the need for a reference to guide the reader wishing to trace the requirement back to its source.

A set of test data for the low-level (L-) functions based on this document is under development and is intended for use by developers when implementing the evaluation. The data is currently divided into *valid* (expected to convert without error) and *invalid* (expected to give an error). These can be conveniently viewed in PDF form[1] at valid-domains and invalid-domains. The authoritative original data is available in UTF-8 text files valid-domains.txt and invalid-domains.txt.

Low-level functions

For current data on UNASSIGNED, DISALLOWED, CONTEXTJ and CONTEXTO, see this IANA page.

A comprehensive set of test data for these functions is available from the Unicode Consortium. As befits a comprehensive test set, it contains a large number of tests probing for implementation weaknesses in the more obscure areas of the standards, and so lacks the balance of tests required. However, it provides a fruitful source of raw test data.

For these tests, labels complying with Bidi Rule (RFC5893) should be evaluated when relevant.

L-U2A: IDNA2008 - Convert Unicode domain name to ASCII lookup form

Convert a domain name in Unicode to ASCII using the process described in RFC5891 for domain name lookup.

General tests:

[1] PDF is used to display test data due to problems correctly displaying Unicode from the supplementary multilingual plane in Google Docs.

Test ID	Input: domain comprising the following, with expected ASCII output	Expected error	Test purpose	Reference
L-U2AG1	Plain ASCII	None	Verify that ASCII is passed through unaltered	RFC5891
L-U2AG2	Plain ASCII with >3 char TLD	None	Verify long TLDs are handled	RFC5891
L-U2AG3	Permitted non-ASCII from Unicode base multilingual plane with ASCII TLD	None	Verify basic Unicode support	RFC5891
L-U2AG4	Permitted non-ASCII TLD from Unicode base multilingual plane with ASCII rest of domain	None	Verify basic Unicode support	RFC5891
L-U2AG5	Permitted non-ASCII from Unicode base multilingual plane - entire domain	None	Verify basic Unicode support	RFC5891
L-U2AG6	Permitted non-ASCII from Unicode supplementary multilingual plane - entire domain	None	Verify Unicode support for higher planes	RFC5891

Specific tests:

Test ID	Input: domain comprising the following, with expected ASCII output	Expected error	Test purpose	Reference
L-U2AS1	Permitted non-ASCII from Unicode base multilingual plane, labels separated with . FULL STOP (U+002E)	None	Verify basic Unicode support	UTS#46
L-U2AS2	Permitted non-ASCII from Unicode base multilingual plane, labels separated with ．FULLWIDTH FULL STOP (U+FF0E)	None	Verify basic Unicode support	UTS#46
L-U2AS3	Permitted non-ASCII from Unicode base multilingual plane, labels separated with 。IDEOGRAPHIC FULL STOP (U+3002)	None	Verify basic Unicode support	UTS#46
L-U2AS4	Permitted non-ASCII from Unicode base multilingual plane, labels separated with 。HALFWIDTH IDEOGRAPHIC FULL STOP (U+FF61)	None	Verify basic Unicode support	UTS#46
L-U2AS5	Permitted non-ASCII from Unicode base multilingual plane with '–' (two consecutive hyphens) in the third and fourth character positions	Reject	Ensure malformed Unicode is rejected	RFC5891 § 5.4
L-U2AS6	Permitted non-ASCII from Unicode base multilingual plane with a combining mark as a first character	Reject	Ensure malformed Unicode is rejected	RFC5891 § 5.4

L-U2AS7	Permitted non-ASCII from Unicode base multilingual plane but containing a DISALLOWED character in a label	Reject	Ensure malformed Unicode is rejected	RFC5891 § 5.4
L-U2AS8	Permitted non-ASCII from Unicode base multilingual plane but containing a conforming CONTEXTJ character in a label	None	Verify CONTEXTJ support	RFC5891 § 5.4
L-U2AS9	Permitted non-ASCII from Unicode base multilingual plane but containing a non-conforming CONTEXTJ character in a label	Reject	Verify CONTEXTJ support	RFC5891 § 5.4
L-U2AS10	Permitted non-ASCII from Unicode base multilingual plane but containing a conforming CONTEXTO character in a label	None	Verify CONTEXTO support	RFC5891 § 5.4
L-U2AS11	Permitted non-ASCII from Unicode base multilingual plane but containing an UNASSIGNED character in a label	Reject	Ensure malformed Unicode is rejected	RFC5891 § 5.4
L-U2AS12	Permitted non-ASCII from Unicode base multilingual plane but containing a label that is 64 characters or longer in ACE form	Reject	Ensure malformed Unicode is rejected	RFC5891
L-U2AS13	Permitted non-ASCII from Unicode supplementary multilingual plane but containing a DISALLOWED character in a label	Reject	Ensure malformed Unicode is rejected	RFC5891
L-U2AS14	Permitted non-ASCII from Unicode supplementary multilingual plane but containing an UNASSIGNED character in a label	Reject	Ensure malformed Unicode is rejected	RFC5891

L-U2AS15	Permitted non-ASCII from Unicode base multilingual plane not compliant with the requirements for right-to-left characters specified in the Bidi document (RFC5893)	Reject (SHOULD)	See if Bidi checking happens	RFC5891 § 5.4

L-A2U: IDNA2008 - Convert ASCII domain name to Unicode

Convert a domain name in ASCII to Unicode using the process described in RFC5891.

General tests:

Test ID	Input: domain comprising the following ASCII, with expected Unicode output	Expected error	Test purpose	Reference
L-A2UG1	Plain ASCII	None	Verify that ASCII is passed through unaltered	RFC5891
L-A2UG2	Plain ASCII with >3 char TLD	None	Verify long TLDs are handled	RFC5891
L-A2UG3	ACE domain with ASCII TLD	None	Verify basic Unicode support	RFC5891
L-A2UG4	ACE TLD with ASCII rest of domain	None	Verify basic Unicode support	RFC5891
L-A2UG5	Permitted non-ASCII from Unicode base multilingual plane - entire domain	None	Verify basic Unicode support	RFC5891

L-A2UG6	Permitted non-ASCII from Unicode supplementary multilingual plane - entire domain	None	Verify basic Unicode support	RFC5891

Specific tests:

Test ID	Input: domain comprising the following ASCII, with expected Unicode output	Expected error	Test purpose	Reference
L-A2US1	A-label, not all in lowercase	Accept	A-label validation	RFC3492 § 5
L-A2US2	A-label that ends with '-' (hyphen)	Reject	A-label validation	RFC3492 § 5
L-A2US3	A-label that starts with '-' (hyphen)	Reject	A-label validation	RFC3492 § 5

High-level functions

For these tests, labels complying with Bidi Rule (RFC5893) should be evaluated when relevant.

H-DNS: Domain name - syntactic check

Perform a syntactic check on a domain name. Determine whether the name appears to be correctly formed. If any part of the name already appears to be in ASCII form (an A-label), verify it can be converted to Unicode.

This test should run all the tests described in L-U2A: IDNA2008 - Convert Unicode domain name to ASCII lookup form above and verify that the conversion does not produce an error. In addition, the following tests should also be run. These are all *specific* tests.

Test ID	Input: domain comprising the following	Expected error	Test purpose	Reference
H-DNSS1	Permitted non-ASCII from Unicode base multilingual plane with ASCII '.invalid' TLD	None	Verify Unicode support	RFC5891
H-DNSS2	Permitted non-ASCII from Unicode base multilingual plane with empty label ('..')	Reject	Check domain composition	RFC1035
H-DNSS3	Permitted non-ASCII from Unicode base multilingual plane with no label separator character, i.e. none of the following: . FULL STOP (U+002E) ． FULLWIDTH FULL STOP (U+FF0E) 。 IDEOGRAPHIC FULL STOP (U+3002) ｡ HALFWIDTH IDEOGRAPHIC FULL STOP (U+FF61)	Reject	Check domain composition	SAC053

H-ES: Email address - syntactic check

Perform a syntactic check on an email address. Determine whether the address appears to be correctly formed.

General tests:

The *general* test email addresses should include all domain test cases from the *general* tests from Domain name: syntactic check.

Test ID	Input: email address comprising the following	Expected error	Test purpose	Reference

H-ESG1	Plain ASCII local part, '@' permitted non-ASCII from Unicode base multilingual plane domain	None	Verify Unicode support	RFC6531
H-ESG2	Unicode local part from base multilingual plane, '@' plain ASCII domain	None	Verify Unicode support	RFC6531
H-ESG3	Unicode local part from base multilingual plane, '@' permitted non-ASCII from Unicode base multilingual plane domain	None	Verify Unicode support	RFC6531
H-ESG4	Unicode local part from supplementary multilingual plane, '@' permitted non-ASCII from Unicode supplementary multilingual plane domain	None	Verifying local part handling	RFC6531

Specific tests:

Test ID	Input: email address comprising the following	Expected error	Test purpose	Reference
H-ESS1	Plain ASCII local part including '@', '@' plain ASCII domain	Reject	Verifying local part handling	RFC6531
H-ESS2	Quoted plain ASCII string local part including '@', '@' plain ASCII domain	None	Verifying local part handling	RFC6531
H-ESS3	Unicode local part from base multilingual plane including '@', '@' plain ASCII domain	Reject	Verifying local part handling	RFC6531

H-ESS4	Quoted Unicode string local part from base multilingual plane including '@', '@' plain ASCII domain	None	Verifying local part handling	RFC6531
H-ESS5	Unicode local part from supplementary multilingual plane including '@', '@' permitted non-ASCII from Unicode supplementary multilingual plane domain	Reject	Verifying local part handling	RFC6531
H-ESS6	Quoted Unicode string local part from supplementary multilingual plane, '@' permitted non-ASCII from Unicode supplementary multilingual plane domain	None	Verify Unicode support	RFC6531
H-ESS7	Quoted Unicode string local part from supplementary multilingual plane plus '@', '@' permitted non-ASCII from Unicode supplementary multilingual plane domain	None	Verifying local part handling	RFC6531

H-ID: Identifier - Identifier lookup

Compare the identifier stored in the system against the one used to authenticate by the user.

General tests:

Test ID	Input: Registration username	Input: Login	Expected match	Test purpose	Reference
H-IDG1	Plain ASCII identifier	Same plain ASCII identifier	Yes	Verify basic ASCII support	RFC8264
H-IDG2	Unicode identifier, NFD form	Unicode identifier, NFC form	Yes	Verify Unicode equivalence	RFC8264

H-IDG3	Unicode identifier, NFC form	Unicode identifier, NFD form	Yes	Verifying Unicode support	RFC8264